Towards Scenario-based Safety Validation for Autonomous Trains with Deep Generative Models

Thomas Decker^{1,2} \boxtimes , Ananta R. Bhattarai^{1,3}, and Michael Lebacher¹

¹ Siemens AG, Munich, Germany
² Ludwig Maximilians Universität, Munich, Germany
³ Technical University of Munich, Munich, Germany
{thomas.decker;michael.lebacher}@siemens.com; ananta.bhattarai@tum.de

Abstract. Modern AI techniques open up ever-increasing possibilities for autonomous vehicles, but how to appropriately verify the reliability of such systems remains unclear. A common approach is to conduct safety validation based on a predefined Operational Design Domain (ODD) describing specific conditions under which a system under test is required to operate properly. However, collecting sufficient realistic test cases to ensure comprehensive ODD coverage is challenging. In this paper, we report our practical experiences regarding the utility of data simulation with deep generative models for scenario-based ODD validation. We consider the specific use case of a camera-based rail-scene segmentation system designed to support autonomous train operation. We demonstrate the capabilities of semantically editing railway scenes with deep generative models to make a limited amount of test data more representative. We also show how our approach helps to analyze the degree to which a system complies with typical ODD requirements. Specifically, we focus on evaluating proper operation under different lighting and weather conditions as well as while transitioning between them.

Keywords: Operational Design Domain (ODD) · Safety Validation · Deep Generative Models · Autonomous Train · Rail-Scene Segmentation.

1 Introduction

Artificial Intelligence (AI) enables technologies that can process vast amounts of data from various sources in real time and its potential for autonomous vehicles is progressively transforming the transportation industry. This is especially true for the railway domain, where driverless trains are associated with various economic and societal benefits [14]. Moreover, fully automated trains are already in service for many years in constrained and well-controlled environments such as metro lines with platform screen doors [2]. However, enabling operation in general open settings is significantly more demanding as trains are constantly required to perceive and interact with the current environment. While AI has shown promising capabilities in this regard [12], it is still unclear how to rigorously assure the safety of such systems from a regulatory and legal perspective

2 T. Decker et al.

[2]. A popular approach to conduct safety validation of automated vehicles is scenario-based testing [10]. Ideally, fully automated trains are expected to handle any environmental conditions and even unexpected events in a safe and robust manner, but the resulting space of possible scenarios is infeasible to test globally. As a consequence, scenario-based testing is typically performed considering a predefined Operational Design Domain (ODD) [5] which refers to all specific conditions under which a system is strictly required to behave properly including physical, geographical and regulatory constraints [4]. While there already exist proposals regarding ODD specifications for railway applications [13], collecting sufficient test cases covering all relevant aspects and systematically conducting appropriate evaluations still remains challenging. However, AI-powered data generation in the form of deep generative models has demonstrated remarkable capacities to realistically simulate complex data structures [1]. In this work, we propose a framework to systematically leverage deep generative models for scenario-based testing and summarize our practical experiences. Specifically, we create high-resolution image data with conditional Generative Adversarial Networks (cGANs) [15] allowing us to fix high-level image contents, such as the position of rails or other objects while altering different ODD-related characteristics during simulation. In this way, we can make a limited number of test cases more representative for the purpose of safety validation. We further apply our approach to test a camera-based rail-scene segmentation model that is implemented via a deep neural network [17]. Such systems enable accurate perception of the frontal environment which is crucial for safe train operation and obstacle detection [11]. We demonstrate how to perform a systematic model evaluation under natural perturbations like different lighting and weather conditions as well as while transitioning between them. Such an analysis complements classical robustness certification [9,6] and provides an additional tool to validate system safety in a comprehensive way.

2 Background

GANs Generative adversarial networks (GANs) are a popular category of deep generative models that have been extensively studied in computer vision and demonstrate remarkable capabilities to simulate realistic images and videos [7]. GANs consist of two neural networks, a generator and a discriminator, that are trained in concert to create new samples resembling the training data. Conditional GANs (cGANs) are extended versions that allow controlling the properties of generated data via additional input arguments. For images, cGANs enable semantic editing, style translation or creating images with specific details [8].

Semantic Segmentation and RailSem19 Semantic segmentation describes the task of dividing an image into semantically distinct sub-regions and assigning them a corresponding label. Deep neural networks attain state-of-the-art performances for this purpose and have also been applied in corresponding railway applications [11]. Such models are typically trained via labeled training data



Fig. 1: Proposed approach for scenario-based ODD validation with cGANs

comprising images and matching ground truth semantic label masks. A popular metric to evaluate segmentation performance is the Intersection over Union (IoU) score ranging from 0 to 1, where a score of 1 indicates a perfect match between ground truth and the predicted regions and 0 means no overlap. RailSem19 [16] is a publicly available dataset for semantic segmentation of railway scenes. It contains 8500 high-resolution images of real train and tram front views together with pixel-wise semantic labels corresponding to 19 different classes. The provided labels allow to distinguish a variety of different safety-critical objects such as rails, cars, humans or other on-rail vehicles. The dataset also covers various different operation environments, illuminations and weather conditions which all resemble typical components of ODD descriptions for railway applications [13].

3 Proposed Methodology

The goal of our approach is to leverage deep generative models in a systematic way to validate if an AI-powered model fully complies with specific ODD requirements given only a limited amount of test cases. Our proposed methodology is illustrated in Fig 1. As usual for safety validation, we suppose access to a predefined ODD description as well as a set of representative scenario data (a). In our use case, an ODD might among other things also require models to work well under changing lighting conditions and the extensive RailSem19 dataset provides corresponding scenarios. As a second step, we utilize the scenario data for training a cGAN to enable conditional generation of new relevant scenarios (b). In particular, we choose the pix2pixHD architecture [15] that enables the creation of high-resolution images via a generator G receiving two distinct inputs. First, the semantic structure of the desired image can be controlled by providing a semantic label mask s informing G where in the image specific objects or structures should appear. Second, a separate encoder network E was



Fig. 2: Styles represented by cluster centers of class Sky: cloudy, sunny and night.



Fig. 3: Synthesizing snowfall by altering features of different semantic categories.

designed to grasp the stylistic characteristics of different semantic categories. More precisely, E encodes low-level details of regions in x into low-dimensional feature vectors z forming a numerical style space. This setup allows us to semantically manipulate a given scenario to increase test capacities and improve ODD coverage. To do so, we first run the trained encoder on all instances in the training set and save the resulting feature vectors. Following [15], we perform clustering on these features for each semantic category to localize ODD-related concepts in the style space (c). For instance, the cluster centers for the category Sky can encode styles such as sunshine, cloudiness or night. This enables us to synthesize new realistic images with identical high-level structures determined by s but exhibiting different stylistic properties, like the same railway scene under varying lighting conditions. Moreover, we can also simulate continuous transitioning between two styles by interpolating the corresponding style encodings during image generation. To systematically test how well a model complies with an ODD requirement we can semantically manipulate available scenarios to exhibit specific properties and evaluate its effect on the model's performance (d). In the case of rail-scene segmentation this methodology allows us for instance to explicitly validate if a model works sufficiently well under sunny, cloudy or nighttime illumination.

4 Results and Experiences

Scenario Simulation To simulate test scenarios we trained the proposed cGAN on RailSem19 based on the default implementation provided by the authors [15].



Fig. 4: **Top**: Original image, and synthesized versions with minor artifacts. **Bottom**: Original image, and synthesized versions with significant artifacts.

Applying k-means clustering to all style encoding indeed enables us to locate k=10 distinct regions in the style space that correspond to different lighting and weather conditions. Fig. 2 shows some styles represented by cluster centers for the sky class, which we refer to as prototypical cloudy, sunny, and night during our experiments. To manipulate illumination, we replace the feature vector of the sky instance in a given image with desired cluster center and synthesize a new image as outlined in Section 3. Changing the weather to snowfall involves manipulating several instances in the railway scene individually. Therefore, we replace the original style features of all semantic classes with their respective cluster centers that best depict the snowfall weather condition. Fig. 3 shows how the features of each semantic category are altered to translate an original weather condition into snowfall. Overall, images with a significant amount of sky, vegetation, terrain or rails are of high quality, e.g. Fig.4. However, we also observed significant artifacts while encoding images with buildings, people and cars. Fig. 4 shows such an example where also the simulation of snow fails. Hence, model evaluations on synthesized examples should ideally be complemented by manual human inspection on a case-by-case basis to ensure sound conclusions.

Model Evaluation For our experiments we use the PSPNet [17], which we train on RailSem19 similarly to the procedure in [16]. Out of the 8500 available images we randomly selected 7140 for training and fine-tuning leaving us only

6 T. Decker et al.

1360 for rigorous testing. On this test set, we achieve a mean IoU of 0.65 over all classes which is comparable with the reference performance reported in [16]. To validate if the model also complies with the ODD-related requirements of proper operation under different lighting conditions and snow we applied our proposed methodology to create 4 new versions of the original test set where we modified the style of all images accordingly. The corresponding IoU scores per segmentation class are reported in Fig. 5. Our evaluation reveals that in all scenarios the model performs well with respect to the detection of tram/rail tracks or trackbeds but seems to struggle with traffic lights/signs or trucks. Also, simulating nighttime conditions seems to be particularly detrimental to the model performance, as for instance indicated by the significant IoU drops for segmenting cars, humans, construction sites or other on-rail vehicles. Since accurate detection of corresponding objects is potentially safety-critical our evaluation possibly reveals a crucial deficiency. To verify if this is indeed the case or just due to simulation artifacts we can also evaluate the model behavior on individual examples transitioning from their original style to night mode. Fig. 6 displays an image with an on-rail vehicle in front of the train, that is accurately detected under the original illumination. Progressively moving to night causes the model to miss the object, but its visual appearance also becomes unnatural requiring closer inspection by a human auditor. Moreover, by evaluating other images under style transition we can demonstrate other deficiencies. In Fig. 7 the model performs well on the original image. Since it is already sunny, the synthesized sunny version is quite similar but the rail tracks are perceived as tram tracks by the model. Surprisingly, when transitioning towards night conditions the prediction suddenly turns correct at some point, although the visual appearance of the rails changes only marginally. Similarly in Fig. 8, moving to snow causes the model to suddenly confuse rail and tram tracks despite the high visual similarity of the tracks in all pictures.

5 Conclusion

In this work we report our experiences with cGANs to validate if an AI-powered model complies with typical ODD requirements, especially varying weather and lighting conditions. We intend to expand the approach to also enable the rendering of new objects such as obstacles, persons or vehicles on the rails. Comparing the simulation quality of similar generative model types, such as variants of recently popularized Diffusion Models [3] is also relevant for future work.

Acknowledgement We acknowledge the support from the Federal Ministry for Economic Affairs and Climate Action (BMWK) via grant agreement 19I21039A.

References

1. Bond-Taylor, S., Leach, A., Long, Y., Willcocks, C.G.: Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autore-

	Road	Side- walk	Con- struction	Tram- track	Fence	Pole	Traffic- light	Traffic- sign	Vege- tation	Terrain	Sky	Human	Rail- track	Car	Truck	Trackbed	On- rails	Rail- raised	Rail- embed.
Original	0.59	0.58	0.73	0.74	0.52	0.56	0.40	0.43	0.86	0.67	0.95	0.62	0.89	0.75	0.34	0.74	0.73	0.68	0.55
Cloudy	0.54	0.53	0.68	0.75	0.47	0.53	0.33	0.25	0.82	0.63	0.96	0.50	0.90	0.68	0.13	0.72	0.63	0.70	0.56
Sunny	0.53	0.53	0.68	0.73	0.48	0.55	0.36	0.27	0.82	0.61	0.96	0.48	0.89	0.66	0.11	0.72	0.58	0.69	0.53
Night	0.40	0.42	0.29	0.59	0.25	0.32	0.04	0.07	0.52	0.46	0.14	0.32	0.84	0.47	0.04	0.65	0.17	0.61	0.45
Snow	0.45	0.43	0.61	0.64	0.40	0.49	0.35		0.78	0.49	0.95	0.47	0.86	0.63	0.09	0.55	0.56	0.64	0.51

Fig. 5: Class-wise IoU results of the trained segmentation model on test data



Fig. 6: Change in IoU for an on-rail object when changing the original lighting to night-time. Huge performance decline when going from 6b to 6c.



Fig. 7: Change in IoU of rail tracks when moving from original to night illumination. Unstable performance when going from 7a to 7b and 7c to 7d.



Fig. 8: Change in IoU of the rail tracks when changing the original weather condition to snow. Despite similarity, performance drops from 8b to 8c.

8 T. Decker et al.

gressive models. IEEE transactions on pattern analysis and machine intelligence (2021)

- Flammini, F., De Donato, L., Fantechi, A., Vittorini, V.: A vision of intelligent train control. In: Reliability, Safety, and Security of Railway Systems. Modelling, Analysis, Verification, and Certification: 4th International Conference, RSSRail 2022, Paris, France, June 1–2, 2022, Proceedings. pp. 192–208. Springer (2022)
- Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. arXiv preprint arXiv:2210.09276 (2022)
- 4. Koopman, P., Fratrik, F.: How many operational design domains, objects, and events? Safeai@ aaai 4 (2019)
- 5. Koopman, P., Wagner, M.: Toward a framework for highly automated vehicle safety validation. SAE Technical Paper, Tech. Rep (2018)
- Li, L., Xie, T., Li, B.: Sok: Certified robustness for deep neural networks. In: 44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, 22-26 May 2023. IEEE (2023)
- Liu, M.Y., Huang, X., Yu, J., Wang, T.C., Mallya, A.: Generative adversarial networks for image and video synthesis: Algorithms and applications. Proceedings of the IEEE 109(5), 839–862 (2021)
- Pang, Y., Lin, J., Qin, T., Chen, Z.: Image-to-image translation: Methods and applications. IEEE Transactions on Multimedia 24, 3859–3881 (2021)
- Paterson, C., Wu, H., Grese, J., Calinescu, R., Păsăreanu, C.S., Barrett, C.: Deepcert: Verification of contextually relevant robustness for neural network image classifiers. In: Computer Safety, Reliability, and Security: 40th International Conference, SAFECOMP 2021, York, UK, September 8–10, 2021, Proceedings 40. pp. 3–17. Springer (2021)
- Riedmaier, S., Ponn, T., Ludwig, D., Schick, B., Diermeyer, F.: Survey on scenariobased safety assessment of automated vehicles. IEEE access 8, 87456–87477 (2020)
- Ristić-Durrant, D., Franke, M., Michels, K.: A review of vision-based on-board obstacle detection and distance estimation in railways. Sensors 21(10), 3452 (2021)
- Tang, R., De Donato, L., Besinović, N., Flammini, F., Goverde, R.M., Lin, Z., Liu, R., Tang, T., Vittorini, V., Wang, Z.: A literature review of artificial intelligence applications in railway systems. Transportation Research Part C: Emerging Technologies 140, 103679 (2022)
- Tonk, A., Boussif, A., Beugin, J., Collart-Dutilleul, S.: Towards a specified operational design domain for a safe remote driving of trains. In: Proceedings of the 31st European Safety and Reliability Conference, Angers, France. pp. 19–23 (2021)
- Trentesaux, D., Dahyot, R., Ouedraogo, A., Arenas, D., Lefebvre, S., Schön, W., Lussier, B., Chéritel, H.: The autonomous train. In: 2018 13th Annual Conference on System of Systems Engineering (SoSE). pp. 514–520. IEEE (2018)
- Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: Highresolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
- Zendel, O., Murschitz, M., Zeilinger, M., Steininger, D., Abbasi, S., Beleznai, C.: Railsem19: A dataset for semantic rail scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
- Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)