Generation of road zone synthetic data for training MOT models with the NVIDIA Omniverse platform

David Conde, Joaquín Martínez, Jesús Balado, Pedro Arias
GeoTECH, CINTECX, Universidade de Vigo,Vigo, Spain
david.conde.morales@uvigo.gal

Abstract. Using synthetic data alongside real data is a powerful technique for enhancing the performance of machine learning models. To this end, a new tool based on the NVIDIA Omniverse platform has been developed to generate synthetic data that closely mimics real-world examples. The use of synthetic data can mitigate problems such as class imbalance and data sparsity, and help prevent overfitting to the training set. By creating additional data points that resemble underrepresented classes, synthetic data can balance the data distribution and reduce differences between classes, resulting in better generalization performance and more accurate predictions on unseen data. In general, incorporating synthetic data into the training process can significantly improve the performance and robustness of machine learning models, especially when working with complex real-world datasets.

1. Introduction

In many computer vision applications, deep learning-based approaches have surpassed the use of traditional image processing procedures. However, nowadays most models available for these tasks require large amounts of data for training. The acquisition of data and their corresponding annotations is both time and cost consuming, especially for those specific cases where public datasets availability is limited. When deciding for a dataset to be used during training, its usefulness can be affected by the imbalance of classes, classes mismatch, or the lack of resemblance in the images' attributes to the target domain of the case study. In such cases where popular public datasets cannot be used, computer generated synthetic images can partially or totally reduce the need for real images capture and annotation.

Synthetic image generation is a technique that creates realistic images from computer graphics, using the images rendered from assets such as 3D models, textures, lighting, and camera parameters. Generating synthetic images that are indistinguishable from real ones is challenging and requires careful design and evaluation of the graphic engine and the rendering pipeline. An important advantage of this approach to obtain realistic images is counting with the possibility of generating pixel-perfect annotations as well for all the dataset without human intervention (Nikolenko, 2019). Furthermore, some of the most typical annotations such as semantic and instance segmentation, 2D and 3D bounding boxes, or depth and normal maps, can be directly extracted from the arbitrary output variables (AOV) that the rendering pipeline uses to combine additional information per pixel into the final image. The use of AOV therefore makes the processing overhead not to significantly increase, making the cost of the annotation to be computationally cheap.

Since a perfect graphic recreation of the real scene of interest is limited both technically and by the availability of high-quality assets, domain adaptation helps to align the disparity between domains such that a computer vision model trained with non-ideal source domain data can be generalized to perform its task in a target domain of interest (Farahani et al., 2021). However, this approach is limited by the availability of samples from a target domain, therefore reducing the immediate applicability of generated synthetic data for training.

In this work, a solution based on the NVIDIA Omniverse platform and Pixar's Universal Scene Description (USD) is proposed to ease the acquisition process and require a lesser amount of data. This methodology aims to easily and directly generate annotated photorealistic synthetic data intended to train Multiple Object Tracking (MOT) models for road environments. Generated data is also used in combination of real data for fine-tuning to assess the validity of synthetic images both as the sole source and as a pretraining source. The synthetic data generator is built and validated in a scenario of road traffic monitoring from a drone, due to the lack of public datasets with similar enough features.

2. Related Work

Synthetic data generation of images is a rising trend in computer vision due to a variety of key features. For instance, allows for the generation of images in large quantities and with high diversity, while keeping the data tailored to specific tasks and scenarios. In addition, the produced images can be accurate visualizations of scenes with complex and realistic physics, lighting, occlusion, and noise, which can enhance the robustness and accuracy of models.

A direct approach to take advantage of computer-generated images is using existent interactive software with visually similar scenes to the domain of interest in a case study. Such is the case of the GTA5 dataset (Richter et al., 2016), that uses a driving focused video game to capture data rendered with the respective AOVs to write the automatic pixel-perfect annotations. By capturing data visually alike the CityScapes dataset (Cordts et al., 2016) and with a compatible annotation format, they allow to effectively enlarge the total amount of data to use in a computer vision model targeting to a domain with analogous features. In a similar fashion, Sims4Action (Roitberg et al., 2021) employs a social simulator game to build a dataset intended to train models for human actions recognition.

While the use of video games or other interactive content offer an immediate data extraction provided the pertinent tools, these are not always available. In many cases accessing the AOVs requires the modification of program files or use of graphics debugging tools. Moreover, the obtainable data is usually somewhat limited by the original creators' intention, not proving useful for broader applications and being subject to legal issues for the use of their assets without license. Free to use graphic engines like Unity or Unreal Engine offer however a lower-level basis over which developers can build their own environments without being subject to the aforementioned restrictions of a final commercial application. The SYNTHIA dataset (Ros et al., 2016) is composed by renders of city landscapes captured from Unity with their annotations.

With game engines not being originally aimed for synthetic data generation and in some cases being source closed, creation of annotation data would require writing custom shaders instead of taking advantage of the non-accessible AOVs. Solutions in the form of plugins like the NVIDIA Deep learning Dataset Synthesizer (NDDS) (To et al., 2018) for Unreal Engine 4 ease the task of retrieving annotations and frames from the renderer under specific user directives in the environment. More recently, Unity Technologies released Unity Perception (Borkman et al., 2021) for performing similar functions, claiming to be able to achieve better accuracy with the combination of synthetic and real data than using only the later.

Synthetic datasets have served as well to evaluate new techniques for adversarial domain adaptation. For instance, GTA5 and SYNTHIA has been used to validate CyCADA (Hoffman et al., 2018), that performs cycle-consistent domain adaptations using generative adversarial networks.

In the later years, NVIDIA released their new Omniverse software, a collaborative platform leveraged by Pixar's USD (Miller et al., 2022) as the primary and extensible data interchange format capable of storing and authoring data containing geometries, materials, physics, and behaviors among others. USD objects can as well define variant sets to specify different attribute variations on the same object type, instancing in memory only once the common elements needed to render the scene. These features, in addition to the NVIDIA Omniverse Replicator tool to ease synthetic data generation (SDG), have been used to promote the platform as a new standard for SDG.

This work will use the newer NVIDIA Omniverse platform to build an Omniverse extension where the user can parametrize the most common properties and generate annotated synthetic data with little effort. The quality of the generated data will be validated with a YOLOv8 detection model using real life images of live traffic captured from a UAV.

3. Methodology

The development of this work will follow the design and implementation of an Omniverse extension capable of generating general purpose datasets with specified 3D assets. It will be used to parametrize the dataset employed to validate the proposed approach in a visual detection task.

3.1 Omniverse Extension

NVIDIA Omniverse is a modular computer graphics oriented platform with emphasis on collaborative workflows for creative and industrial applications. One of the core components of the provided toolset is the Replicator SDK, which provides functionality for easily scripting the generation of physically accurate 3D synthetic data already annotated for training and testing AI perception models. Replicator can be used to create synthetic data for various perception tasks, such as object detection, segmentation, pose estimation, and depth estimation.

To facilitate the creation of custom datasets without the need to write code, in this work is presented an extension for Omniverse that allows the user to interact with Replicator with a UI as shown in Figure 1.

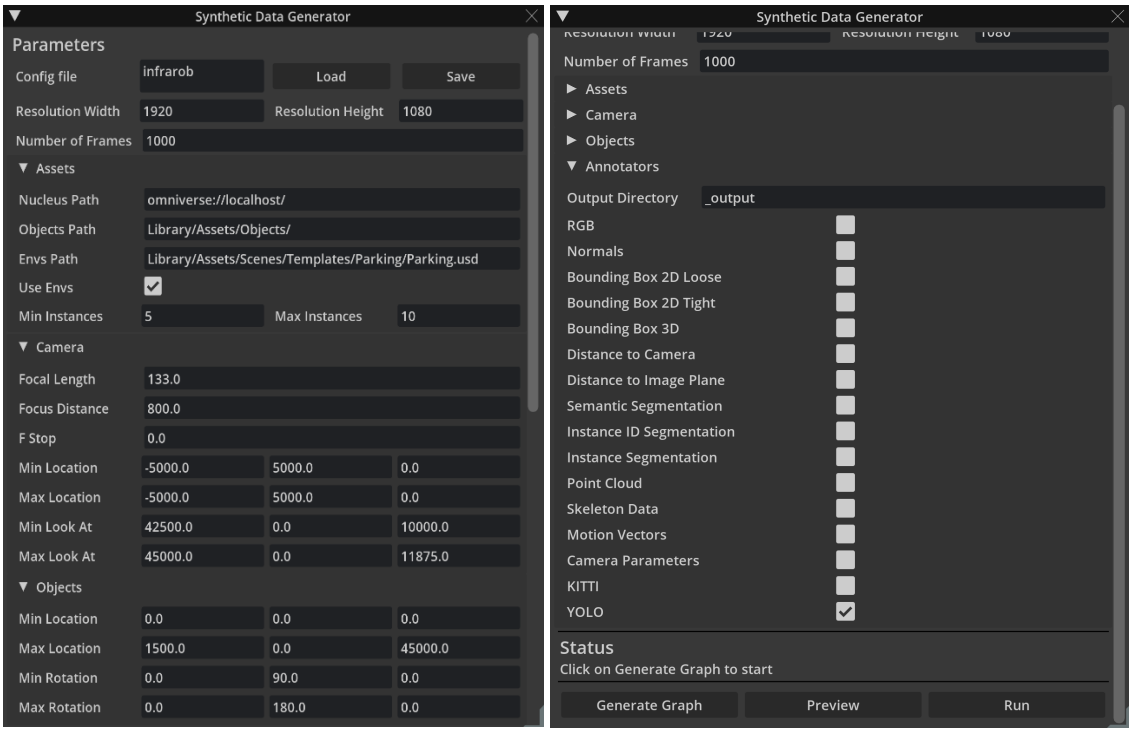

Figure 1: General settings of the Omniverse extension

The displayed extension allows to modify camera attributes, define posing values for objects to be instantiated and define paths to look for the assets to be used with online generated semantic data. It also includes extra functionality to support storing and loading experiments parameters, as well as converting between annotation formats. While by default Replicator stores annotations as Python dictionaries with references to NumPy objects, this extension accommodates the format conversion to other popular data layouts as the one used by YOLO.

3.2 Dataset Parametrization

For comparability with real data, the synthetic dataset was parametrized to resemble as closely as possible to a generalized version of a current real dataset being captured with images from a UAV with large focal lens cameras flying over 50 meters above a live traffic area and looking at further 450 meters. These images were taken with a DJI Matrice 300 RTK and a Zenmuse H20 camera to be later annotated with the first 8 classes of the COCO dataset, namely being: person, bicycle, car, motorcycle, airplane, bus, train, and truck.

To optimize different environments loading and facilitate acquisition, HDR skyboxes are used instead of actual environment models, with farthest vehicles being rendered as smaller objects compared to the scene, but effectively providing small screen space scales suitable for the AI training. Figure 2 shows an example of the generated dataset with segmentation and bounding box annotations.

Figure 2: Example of automatically generated annotations from AOVs

3.3 Detection AI Training

To assess the quality of the synthetic dataset for training purposes, training experiments have been conducted on non-pretrained models and validated on a split of the same target domain with the following considerations:

- Using just a smaller training split of the real dataset with 98 images for training.
- Using the full real dataset with 980 images.
- Using 1000 images of the synthetic dataset, fine-tuned with the previous 98 real images.

The carried-out tests involve the analysis of Precision-Recall curves for each one of the independent classes and their average to identify the contribution of synthetic data on the final model. This evaluation requires a small dataset with annotated real data as well as a more extensive synthetic dataset built with the described tool, parametrizing it to achieve the generation of images alike the target domain where the tests will be carried out.

4. Results

The aforementioned criteria for running training experiments were applied beginning with the training of just 10% of the full real dataset, consisting of 98 images. Figure 3 shows the Precision-Recall curve for the validation of this training, displaying the average precision per class, as well as the mean for all of them.

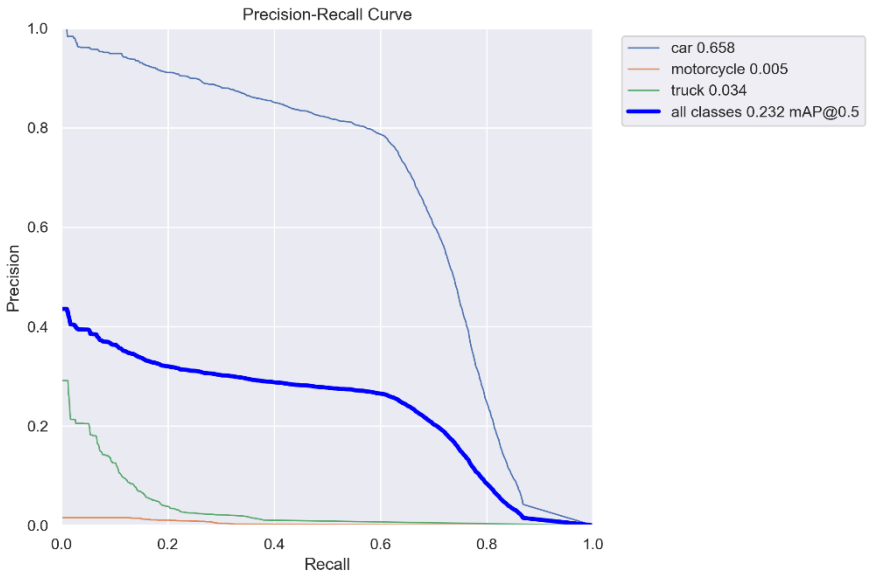

Figure 3: Precision-Recall curve when training with few real data (98 images)

A highlightable insight into the individual class's curves is the large variance among cars and other types of vehicles. This is due to the fact that the presence of cars in the real dataset is predominant among all types of vehicles, with motorcycles being far less abundant on highways and also harder to detect by their smaller size.

As could be expected, since the initial experiment ran on too few images, increasing the training dataset to cover all the captured 980 images leads to a vast increase in the model performance as shown in Figure 4. It is important to note that the mean average precision is heavily affected by the inability of the model to detect motorcycle when they are present in the image.

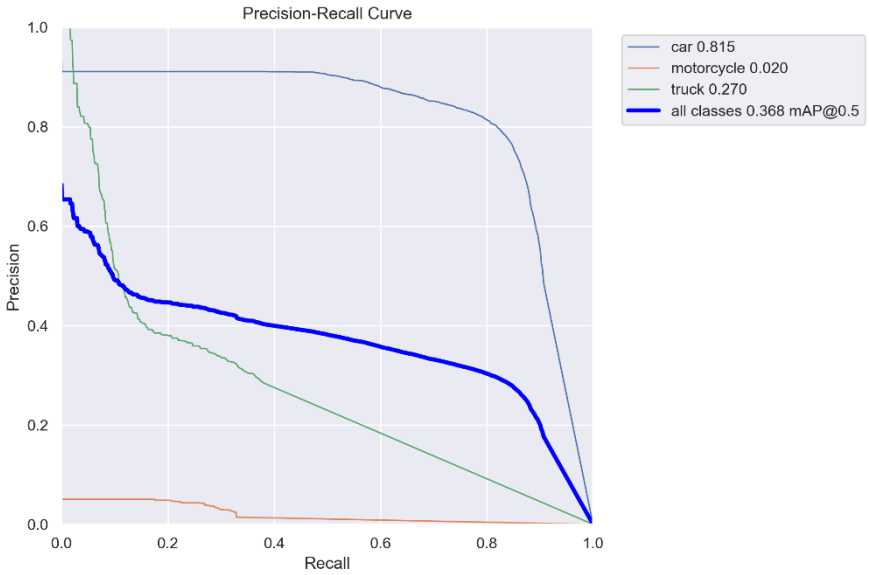

Figure 4: Precision-Recall curve when training with all real data (980 images)

Afterwards, the model was once again trained from scratch just with synthetic data. Even if the generated images are more consistent with inter-class variance, the lack of resemblance to the

target domain in comparison to the real data heavily penalizes the quality of the model. Hence, the more extended approaches consist of:

- Either improve the quality of the dataset by using better assets and composition, or implementing domain adaptation, requiring real samples from the target domain.
- Or using some real samples to fine-tune the model.

After using the small 98 images split from the real dataset, the results depicted in Figure 5 were obtained.

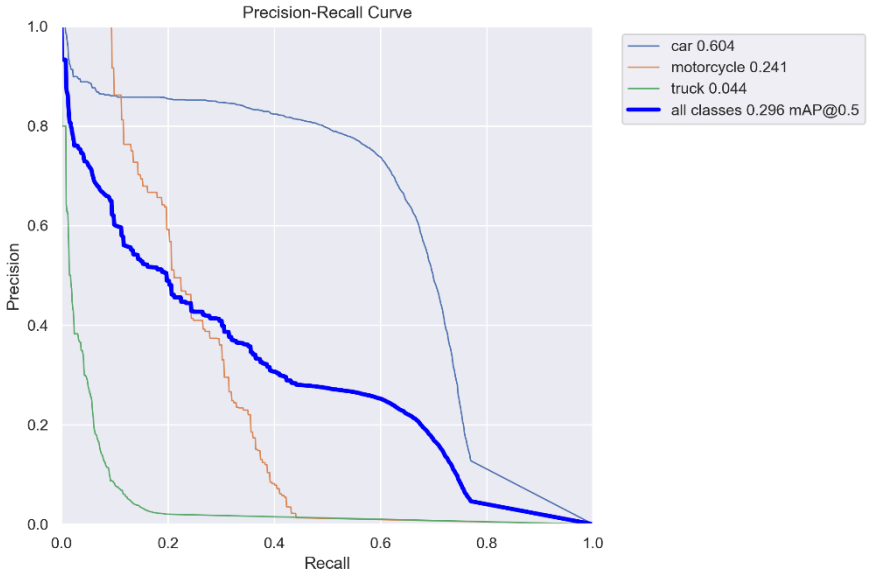

Figure 5: Precision-Recall curve when training with synthetic data (1000 images) and fine-tuned with few real data (98 images)

As can be seen, leveraging the availability of few real data samples with synthetic data helped in this experiment to raise the mean average precision, as well as reducing the inter-class variance by being able to detect the originally more complicated motorcycle class.

5. Conclusions

This work presented a methodology to create and validate synthetic datasets on real scenarios, as well as putting on focus the improvement that a similar approach can suppose when the availability of real data is limited.

The tool for creating synthetic data was implemented in the NVIDIA Omniverse platform as a new extension that will be available both in its current form to integrate in Omniverse applications, as well as in the form of a standalone app running on the same platform.

It was assessed the importance of disposing of high-quality assets in substitution of more real annotated images, or well being able to conduct domain adaptation. Regarding virtual scenes composition, it is proven that even when disposing of few real data, being able to equalize different classes' objects instantiation help to reduce the average precision inter-class variance.

Future lines of work will focus on the public availability of the tool with improvements on its design and the integration of domain adaptation techniques similar to CyCADA to help build and compare new synthetic datasets.

Acknowledgments

This research has received funding from Xunta de Galicia through human resources grant (ED481B-2019-061) and competitive reference group (ED431C 2020/01), from the Government of Spain through project "Software multi-capa para el procesamiento online de datos lidar enfocado a la monitorización del transporte" with reference PDC2022-133851-I00 funded by MCIN/AEI/10.13039/501100011033, and from the European Union's "NextGenerationEU"/PRTR". This paper was carried out in the framework of the InfraROB project (Maintaining integrity, performance and safety of the road infrastructure through autonomous robotized solutions and modularization), which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 95533. It reflects only the authors' views. Neither the European Climate, Infrastructure, and Environment Executive Agency (CINEA) nor the European Commission is in any way responsible for any use that may be made of the information it contains.

References

Borkman, S., Crespi, A., Dhakad, S., Ganguly, S., Hogins, J., Jhang, Y.-C., Kamalzadeh, M., Li, B., Leal, S., Parisi, P., Romero, C., Smith, W., Thaman, A., Warren, S., & Yadav, N. (2021). *Unity Perception: Generate Synthetic Data for Computer Vision*. https://arxiv.org/abs/2107.04259v2

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Farahani, A., Voghoei, S., Rasheed, K., & Arabnia, H. R. (2021). *A Brief Review of Domain Adaptation*. 877–894. https://doi.org/10.1007/978-3-030-71704-9_65/COVER

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A. A., & Darrell, T. (2018). *CyCADA: Cycle-Consistent Adversarial Domain Adaptation* (pp. 1989–1998). PMLR. https://proceedings.mlr.press/v80/hoffman18a.html

Miller, T., Li, H., Karanam, N., Sinno, N., & Scopio, T. (2022). Making Encanto with USD: Rebuilding a Production Pipeline Working from Home. *Proceedings - SIGGRAPH 2022 Talks*. https://doi.org/10.1145/3532836.3536236

Nikolenko, S. I. (2019). Synthetic Data for Deep Learning. *Springer Optimization and Its Applications*, *174*, 1–54. https://doi.org/10.1007/978-3-030-75178-4_1

Richter, S. R., Vineet, V., Roth, S., & Koltun, V. (2016). Playing for Data: Ground Truth from Computer Games. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *European Conference on Computer Vision (ECCV)* (Vol. 9906, pp. 102–118). Springer International Publishing.

Roitberg, A., Schneider, D., Djamal, A., Seibold, C., Reiß, S., & Stiefelhagen, R. (2021). Let's Play for Action: Recognizing Activities of Daily Living by Learning from Life Simulation Video Games. *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., & Lopez, A. M. (2016). *The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes* (pp. 3234–3243).

To, T., Tremblay, J., McKay, D., Yamaguchi, Y., Leung, K., Balanon, A., Cheng, J., Hodge, W., & Birchfield, S. (2018). *NDDS: NVIDIA Deep Learning Dataset Synthesizer*.