

Assessing Simulation-Based Dataset Generation for Object Detection [†]

Richárd Reisinger and János Hollósi *

Vehicle Industry Research Center, Széchenyi István University, 1. Egyetem tér, 9026 Győr, Hungary; reisinger.richard@hallgato.sze.hu

* Correspondence: hollosi.janos@sze.hu

[†] Presented at the Sustainable Mobility and Transportation Symposium 2026, Győr, Hungary, 8 October 2026.

Abstract

Dataset generation is a critical component in training robust perception models for tasks such as object detection and road segmentation, where large-scale annotated data are required. In this work, we present a comparative study of two simulation platforms, CARLA and aiSim, focusing on their effectiveness for synthetic dataset creation. The research evaluates how effectively neural networks trained on synthetic data from each platform can generalize to real-world conditions, with particular emphasis on real-world object detection performance. While both simulators provide automated annotation pipelines, they differ in usability, flexibility, and scalability when constructing labeled datasets. The study specifically focuses on training neural networks using synthetic data from each simulator and evaluating their performance on real-world footage, thereby assessing the domain transfer capability of the generated datasets. This approach allows for a direct comparison of how simulation realism and scenario diversity impact generalization to real-life conditions. The findings provide insights into the limitations of current simulation tools in bridging the gap between synthetic and real-world data and identify the factors that most strongly influence successful transfer learning.

Keywords: dataset generation; object detection; synthetic data; autonomous vehicles; CARLA; aiSim; YOLO; KITTI; domain transfer

1. Introduction

Robust perception for autonomous driving requires substantial amounts of accurately annotated data. For tasks such as object detection and scene understanding, collecting real-world footage and producing high-quality manual annotations is expensive, time-consuming, and difficult to scale. Simulation therefore provides an important alternative by enabling the generation of large synthetic datasets with automatically produced annotations in controlled and repeatable settings.

Simulation also provides researchers with precise control over scene properties, including weather, lighting, traffic density, and object placement, which makes it particularly suitable for generating supervised training data. The key question, however, is how effectively models trained on synthetic data perform under real-world conditions. This synthetic-to-real mismatch, commonly referred to as the domain gap, remains a major challenge in simulation-driven machine learning [1].

Recent research further emphasizes the importance of scalable and diverse simulations for improving robustness and generalization in autonomous driving. Tian et al. [2] introduced SimScale, a scalable simulation framework that synthesizes diverse, high-fidelity



Academic Editors: Gábor Dogossy, Szabolcs Kocsis-Szürke, Dániel Csikor, Dániel Serfőző, Miklós Kuczmann, András Lajos Nagy, László Lendvai and Ambrus Zelei

Published: 15 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

driving states and combines simulated and real-world data through sim-real co-training. Their results demonstrate that increasing simulation data can improve robustness and generalization on challenging real-world benchmarks, highlighting the broader relevance of simulation scale and data diversity to successful sim-to-real transfer.

The present study compares two simulation platforms, CARLA and aiSim, to assess their suitability for synthetic dataset generation in the context of object detection. In addition to considering the practical aspects of dataset creation, the study examines the extent to which neural networks trained on simulator-generated data can transfer successfully to real-world imagery. Particular emphasis is placed on object detection performance under real-world conditions, thereby providing a direct measure of the domain transfer capability of the generated datasets [3,4].

Second, it evaluates all methods under identical experimental conditions, with emphasis on trajectory accuracy, final drift, and runtime.

Third, it identifies practical trade-offs and typical failure modes, including semantic over-masking, dense-mask instability, and sensitivity to stereo depth quality.

2. Materials and Methods

2.1. Simulation Platforms

CARLA is an open-source simulation platform designed for autonomous driving applications. It is widely used in research because of its flexibility, extensive configuration options, and support for automated annotation pipelines. CARLA enables the creation of synthetic urban driving scenarios and provides access to multiple sensors and environmental settings, making it suitable for perception-related experiments [5] (Figure 1).



Figure 1. Example image generated in the CARLA simulation environment used for synthetic dataset creation.

aiSim is a modular, high-reliability automotive simulation environment developed by aiMotive. 5.7.0 version of the software was used for the research. It is intended for verification and validation tasks in the automotive domain and supports both standalone operation and modular SDK-based extension. The architecture of aiSim enables the integration of custom components, including sensors and plugins, and is designed to support reliable simulation workflows in automotive development [6] (Figure 2).

Although both platforms provide automated synthetic data generation, they differ in implementation philosophy, system architecture, and likely visual characteristics. The com-

parison in this paper focuses not only on their software architecture but also on the practical effectiveness of the synthetic data they generate for training object detection models.



Figure 2. Example image generated in the aiSim simulation environment used for synthetic dataset creation.

2.2. Datasets

Two synthetic datasets were generated for this study: one using CARLA and one using aiSim. The datasets were constructed using the same three object categories to support a controlled comparison. However, the resulting datasets were not identical in image or object count, and this difference must be considered when interpreting the transfer results.

Two synthetic datasets were generated for this study: one using CARLA and one using aiSim. To support a fair comparison, the datasets were designed to be similar in size and to contain the same object categories. The selected classes were as follows:

- Car;
- Cyclist;
- Pedestrian.

These categories were chosen because they are among the most relevant and safety-critical objects in road traffic scenes. They are also common categories in both simulation-based training setups and real-world autonomous driving benchmarks.

The aiSim dataset has 5979 images in total: 4784 for training (80.0%) and 1195 for validation (20.0%).

In the training split, 4382 images are non-empty, and 402 are empty. In validation, 1101 are non-empty, and 94 are empty. The training set contains 20,221 labeled objects: 16,415 cars, 178 cyclists, and 3628 pedestrians.

The validation set includes 5108 labeled objects: 4134 cars, 41 cyclists, and 933 pedestrians.

The CARLA dataset contains 4000 images, split into 3200 training images (80.0%) and 800 validation images (20.0%). Within the training subset, 3101 images are non-empty, and 99 are empty, while the validation subset contains 776 non-empty and 24 empty images. The CARLA training set contains 7268 annotated objects, comprising 6253 cars, 299 cyclists, and 716 pedestrians. The validation set contains 1798 annotated objects, including 1551 cars, 64 cyclists, and 183 pedestrians.

aiSim is also more densely labeled, averaging 4.23 objects per training image, versus 2.27 in CARLA. Thus, the aiSim training set contains approximately 49.5% more images and 178.3% more annotated objects than the CARLA training set. These differences are important when interpreting class-specific detection performance and synthetic-to-real transfer results. The present experiments therefore compare the complete datasets gener-

ated under the respective simulation pipelines rather than isolating simulator effects under strictly matched data volume.

aiSim is also more densely labeled, averaging 4.23 objects per training image, versus 2.27 in CARLA.

These statistical differences are important when interpreting class-specific detection performance and synthetic-to-real transfer results (Table 1, Figure 2).

Table 1. Statistical summary of the synthetic datasets.

Dataset	Split	Total Images	Car	Cyclist	Pedestrian	Total Objects
aiSim	Train	4784	16,415	178	3628	20,221
aiSim	Validation	1195	4134	41	933	5108
CARLA	Train	3200	6253	299	716	7268
CARLA	Validation	800	1551	64	183	1798

The dataset statistics show that image count, object count, class distribution, and annotation density differ between the two simulation sources. These factors may independently affect training and transfer performance. Consequently, the observed advantage of aiSim should not be attributed exclusively to simulation realism or platform characteristics; a strictly controlled comparison using matched image and object counts would be required to isolate those effects.

In addition to dataset size, both datasets should be described statistically to support a more complete interpretation of the results. In particular, the manuscript should report the number of labeled instances in each class, the relative frequency of each class, and the average number of annotated objects per image. Such statistics are important because differences in class distribution may directly influence class-wise detection performance and synthetic-to-real transfer.

2.3. Detection Models

To assess the quality and transferability of the generated datasets, four YOLO-based [7,8] object detection models were selected:

- YOLOv8l [9];
- YOLO12L [10];
- YOLO26m;
- YOLO26L [11].

These models were chosen to compare multiple detector variants rather than rely on a single architecture. Each selected model was trained separately on the CARLA- and aiSim-generated datasets. As a result, the experimental setup produced a set of corresponding models whose performance could be directly compared with that of the simulator used to generate the training data.

2.4. Training Configuration

After training, we selected the best-performing checkpoint from each experiment for real-world evaluation. We then tested these models on the KITTI dataset, which was used as the real-world benchmark to assess how well performance generalizes beyond the synthetic domain [12].

The evaluation procedure compared YOLO prediction outputs to KITTI ground-truth annotations. Since the YOLO format stores bounding boxes in normalized coordinates, whereas KITTI uses pixel-space bounding box annotations, the predictions first had to be

converted into the KITTI-compatible coordinate system. After this conversion, predictions were matched to ground-truth objects of the same class using an IoU threshold of 0.5.

The analysis considered the following metrics:

- Intersection over Union (IoU);
- Precision;
- Recall;
- F1-score;
- Mean average precision at 0.5 IoU threshold (mAP@0.5).

Although KITTI contains additional categories beyond the three classes used in training, the interpretation of the results in this paper is restricted to the overlapping classes of car, cyclist, and pedestrian. This restriction is necessary because the trained models were not designed to detect the additional KITTI classes.

3. Results

3.1. Overall Performance Comparison

The experiments showed a clear gap between the two synthetic data sources. In every comparable setup, models trained on aiSim data performed better than those trained on CARLA data when tested on the KITTI benchmark.

Among the aiSim-based models, YOLO12L achieved the highest overall mAP@0.5 of 0.0866, while YOLO26L achieved the highest mean IoU of 0.8003. The YOLO26m aiSim result was available and is included in Table 2, with a mean IoU of 0.7988, an mAP@0.5 of 0.0858, a precision of 0.5710, a recall of 0.3885, and an F1-score of 0.4627. For the CARLA dataset, the highest mAP@0.5 was achieved by YOLOv8l with 0.0204, which nevertheless remained substantially below the performance of the aiSim-trained detectors (Table 2).

Table 2. Overall performance comparison of the evaluated YOLO models trained on synthetic datasets generated with aiSim and CARLA and tested on the KITTI dataset.

Model	Training Data	Mean IoU	mAP@0.5	Precision	Recall	F1-Score
YOLOv8l	aiSim	0.7915	0.0770	0.6154	0.3719	0.4636
YOLO12L	aiSim	0.7972	0.0866	0.5694	0.3900	0.4629
YOLO26m	aiSim	0.7988	0.0858	0.5710	0.3885	0.4627
YOLO26L	aiSim	0.8003	0.0860	0.5683	0.3901	0.4626
YOLOv8l	CARLA	0.7077	0.0204	0.3401	0.0272	0.0504
YOLO12L	CARLA	0.7074	0.0176	0.3303	0.0344	0.0622
YOLO26m	CARLA	0.7058	0.0126	0.3675	0.0258	0.0483
YOLO26L	CARLA	0.7199	0.0105	0.4493	0.0217	0.0415

3.2. Class-Wise Performance

A class-wise examination showed that the car class consistently performed best across both simulators. This was expected, as cars are generally larger, more frequent, and visually more stable objects in road scene datasets than pedestrians or cyclists.

For aiSim-trained models, car detection reached AP@0.5 values between 0.3946 and 0.4608, whereas CARLA-trained models remained between 0.0727 and 0.0909. This large gap indicates that the aiSim-generated data produced object representations that transferred far more effectively to real-world car detection.

The pedestrian class also showed a clear advantage for aiSim. The best pedestrian AP@0.5 among aiSim models was 0.1952, compared to 0.0682 for the best CARLA-based model. In some CARLA-trained cases, pedestrian performance approached zero, indicating

that the trained detectors struggled to generalize to real-world pedestrian appearance and scene context.

The cyclist class was the most difficult category for both simulators. Even the better-performing aiSim-based models achieved very low recall values for this class, although their AP@0.5 values remained higher than those of CARLA-trained models. This suggests that cyclists represent a particularly challenging category in synthetic-to-real transfer, likely due to their smaller size, greater shape variability, and lower prevalence in training scenes (Figure 3).

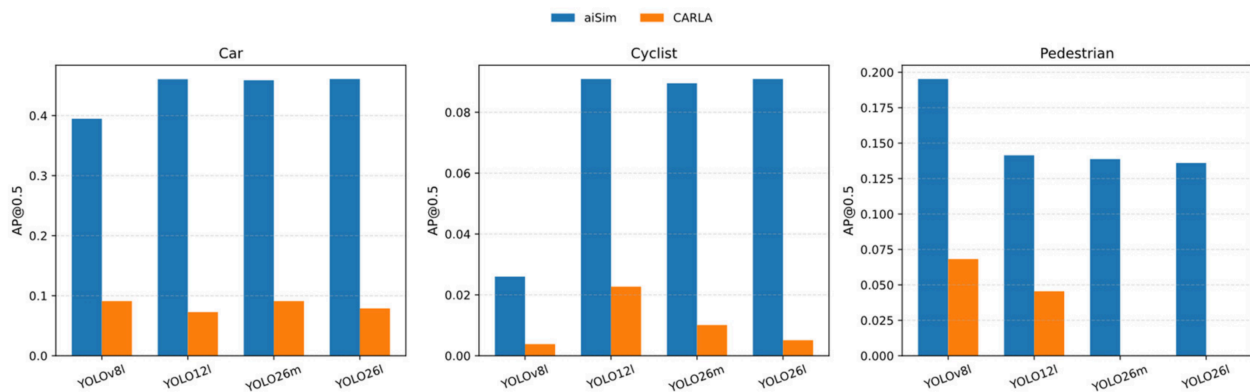


Figure 3. Class-wise AP@0.5 results for YOLOv8l, YOLO12l, YOLO26m, and YOLO26l trained on aiSim- and CARLA-generated datasets and evaluated on the KITTI benchmark.

4. Discussion

The results demonstrate a substantial difference in real-world transfer performance between the two synthetic datasets. Although both CARLA and aiSim were used to create datasets containing the same three object categories, the resulting models showed markedly different performance when applied to real-world imagery. Under the examined training conditions, the aiSim-generated dataset provided a substantially better basis for synthetic-to-real transfer than the CARLA-generated dataset. However, because the datasets differ in both image count and annotation density, the results should be interpreted as a comparison of the complete data-generation pipelines rather than as an isolated measurement of simulator quality.

A likely explanation for the observed performance gap is the combined effect of visual and structural properties of the generated scenes and differences in dataset composition. Transfer to real-world data depends on multiple factors, including rendering realism, texture quality, scene structure, environmental diversity, annotation consistency, class distribution, and the statistical scale of the training data. Recent work such as SimScale [10] likewise highlights the importance of simulation scale and diversity for improving robustness and generalization. In the present study, the higher number of images and annotated objects in the aiSim dataset is an additional factor that may contribute to its stronger performance. Therefore, a matched-data experiment is needed before attributing the full performance difference to simulator-specific characteristics.

A likely reason for this performance gap is the visual and structural quality of the generated scenes. How well a model transfers to real-world data depends on many factors: rendering realism, texture quality, scene structure, environmental diversity, annotation consistency, and the statistical distribution of training examples. So, even when two synthetic datasets look similar in size and class balance, differences in these deeper properties can strongly affect how useful the learned representations are in practice.

The consistently high performance for the car class in both simulators—especially in aiSim—suggests that larger, visually dominant objects benefit the most from synthetic training. Cars usually cover more image area, have more consistent geometry, and appear

often in driving scenes. Those traits make them easier to model in simulation and easier to transfer to real-world data.

In contrast, the poor results for the cyclist class point to an ongoing challenge. Cyclists are typically smaller, appear less often, and vary more in appearance than cars. Their detection is also more sensitive to factors like animation quality, pose variation, and occlusion. The consistently low cyclist recall in both simulators suggests that the synthetic datasets did not capture enough cyclist diversity to generalize reliably to real-world detection.

The pedestrian class occupies an intermediate position in the results. aiSim-trained models achieved meaningful pedestrian detection performance, whereas several CARLA-trained models performed very poorly. This finding may indicate that aiSim provided more transferable visual or geometric representations for human figures under the conditions used in the experiments.

The study also has limitations. First, the analysis was restricted to three classes, which was appropriate for a controlled comparison but does not cover the full range of traffic participants encountered in real-world driving. Second, the aiSim and CARLA datasets differ substantially in image count, annotation count, and annotation density; therefore, the present results do not fully disentangle simulator effects from training-data scale and composition. A follow-up experiment using matched subsets with equal image counts, comparable object counts, and controlled class distributions would provide a stronger basis for isolating simulator-specific effects. Third, the real-world validation relied on a single benchmark dataset, KITTI, which, although widely used, represents only one domain of real-world driving imagery. Future work should therefore include additional benchmarks, more object classes, matched-data comparisons, and potentially domain-adaptation methods or fine-tuning strategies to assess the robustness of simulator-generated training data further.

The study also has limitations. First, the analysis was restricted to three classes, which was appropriate for a controlled comparison but does not cover the full range of traffic participants encountered in real-world driving. Second, the real-world validation relied on a single benchmark dataset, KITTI, which, although widely used, represents only one domain of real-world driving imagery. Future work should therefore include additional benchmarks, more object classes, and potentially domain-adaptation methods or fine-tuning strategies to assess the robustness of simulator-generated training data further.

5. Conclusions

This paper presented a comparative evaluation of CARLA and aiSim as synthetic dataset generation platforms for object detection in autonomous driving. Two datasets were created using the classes car, cyclist, and pedestrian, and four YOLO-based detection models were trained on the generated synthetic data. The trained models were then evaluated on the real-world KITTI dataset using IoU, precision, recall, F1-score, and mAP@0.5.

The results showed a consistent advantage for models trained on aiSim data over those trained on CARLA data in real-world testing. This gap appeared across the key metrics, with especially significant differences in recall and F1-score. Overall, under the tested experimental conditions, the findings indicate that the aiSim-generated dataset provided better transfer to real-world object detection than the CARLA-generated dataset. However, because the aiSim dataset contained substantially more images and annotated objects, the results should not be interpreted as evidence that simulator choice alone caused the observed performance difference.

The results showed a consistent advantage for models trained on aiSim data over those trained on CARLA data in real-world testing. This gap appeared across all key metrics, with especially significant gains in recall and F1-score—two metrics that matter a lot in practical detection settings. Overall, under the tested experimental conditions, the

findings suggest that aiSim produced synthetic training data with clearly better transfer to real-world object detection tasks.

Overall, the study shows that simulation-based dataset generation is a practical approach for developing perception models. At the same time, it demonstrates that transfer performance depends on both the characteristics of the simulator and the scale and composition of the generated dataset.

Future work should broaden the evaluation to larger and more diverse datasets, matched-data experiments with controlled image and object counts, additional model architectures, and domain adaptation methods to better determine how synthetic training pipelines can be optimized for real-world deployment.

Author Contributions: Conceptualization, J.H. and R.R.; methodology, R.R.; software, R.R.; validation, R.R.; formal analysis, R.R.; investigation, R.R.; resources, R.R.; data curation, R.R.; writing—original draft preparation, R.R.; writing—review and editing, J.H.; visualization, R.R.; supervision, J.H.; project administration, J.H.; funding acquisition, J.H. All authors have read and agreed to the published version of the manuscript.

Funding: Project no. TKP2021-NVA-23 has been implemented with the support provided by the Ministry of Technology and Industry of Hungary from the National Research, Development and Innovation Fund, financed under the TKP2021-NVA funding scheme.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data for this study are available upon request from the corresponding author.

Acknowledgments: During the preparation of this manuscript/study, the authors used ChatGPT (OpenAI, GPT-5.5) for the purposes of language editing, stylistic improvement, and formatting assistance. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hu, A.; Murez, Z.; Mohan, N.; Dudas, P.; Hawke, J.; Badrinarayanan, V.; Cipolla, R. Sim-to-Real Domain Adaptation for Lane Detection and Classification in Autonomous Driving. *arXiv* **2022**, arXiv:2202.07133. [CrossRef]
2. Tian, H.; Li, T.; Liu, H.; Yang, J.; Qiu, Y.; Li, G.; Wang, J.; Gao, Y.; Zhang, Z.; Wang, L.; et al. SimScale: Learning to Drive via Real-World Simulation at Scale. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Denver, CO, USA, 3–7 June 2026.
3. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
4. Sadeghi, F.; Levine, S. CAD2RL: Real Single-Image Flight without a Single Real Image. In Proceedings of the Robotics: Science and Systems, Cambridge, MA, USA, 12–16 July 2017.
5. Dosovitskiy, A.; Ros, G.; Codevilla, F.; Lopez, A.; Koltun, V. CARLA: An Open Urban Driving Simulator. In Proceedings of the 1st Annual Conference on Robot Learning, Mountain View, CA, USA, 13–15 November 2017.
6. aiMotive. aiSim. Available online: <https://aimotive.com/aisim/> (accessed on 5 May 2026).
7. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLO. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 10 May 2026).
8. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
9. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8 Documentation, Ultralytics, 2023. Ultralytics Docs. Available online: <https://docs.ultralytics.com> (accessed on 30 May 2026).
10. Ultralytics, YOLO12: Attention-Centric Object Detection, Ultralytics Docs. 2025. Available online: <https://docs.ultralytics.com/models/yolo12> (accessed on 4 June 2026).

11. Ultralytics, Ultralytics YOLO26, Ultralytics Docs. 2026. Available online: <https://docs.ultralytics.com/models/yolo26> (accessed on 4 June 2026).
12. Geiger, A.; Lenz, P.; Urtasun, R. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.