CARLA-Seg

A Synthetic Dataset for Urban Semantic Segmentation

Carlos Ruiz Aguirre (IronDog421)

May 20, 2025

1 Introduction

Training deep learning models for autonomous driving heavily relies on large-scale datasets with precise annotations. However, acquiring such datasets in specific locations or scenarios is often challenging due to the high cost and logistical complexity of data collection and manual annotation. This challenge becomes even more evident in scenarios like urban environments where capturing and annotating real-world images is expensive and sometimes impractical.

To address these limitations, this paper introduces a synthetic dataset generated using the CARLA driving simulator, specifically leveraging the capabilities of the Unreal Engine 5 version and its Upgraded Town 10 environment. The dataset comprises 7,233 training and 1,448 validation items, each containing RGB images at a resolution of 1280×720 pixels, paired with semantic segmentation masks formatted according to YOLOv11 (image masks will also be included in the future with more segmented classes). The annotated classes include standard traffic participants and infrastructure elements crucial for urban driving scenarios: cars, motorcycles, bicycles, traffic lights, buses, pedestrians, roads, sidewalks, and traffic signs.

Our primary motivation is to demonstrate how synthetic datasets can effectively supplement or even replace costly real-world annotations, especially for underrepresented or geographically constrained environments. By employing CARLA's high-fidelity simulation, researchers can generate extensive, diverse, and fully annotated data directly on standard personal computers, significantly lowering the barrier to entry for autonomous driving research.

In this work, we further validate our dataset's utility by integrating it into a broader training regime aimed at improving YOLOv11 model performance for semantic segmentation driving tasks in Valencia—a location where obtaining high-quality annotated data remains prohibitively expensive. Initial results indicate that incorporating this synthetic data significantly enhances the model's understanding of real-world urban scenes.

The dataset is made publicly available on GitHub to encourage widespread adoption and facilitate further research. It is intended to be freely accessible, subject to attribution to the original authors, thereby fostering community-driven advancements in the field of autonomous driving and computer vision.

2 Related Work

2.1 Real-world Driving Datasets

Real-world datasets have significantly advanced autonomous driving research by providing diverse and high-quality annotated data. Prominent examples include Cityscapes, KITTI, Mapillary Vistas, BDD100K, nuScenes, and Waymo Open Dataset. Cityscapes [4] comprises 5,000 finely annotated urban street images, extensively used for benchmarking segmentation models. KITTI [6] is a widely utilized dataset for various tasks including object detection, stereo vision, and semantic segmentation, featuring detailed annotations from real-world driving sequences. Mapillary Vistas [8] and BDD100K [13] expand on these by providing larger-scale datasets (25k and 100k images respectively), covering diverse environmental conditions such as varying weather and lighting. nuScenes [3] and Waymo Open Dataset [11] further include multimodal sensor data (camera, LiDAR, radar), enriching the dataset landscape for robust perception algorithms. Nonetheless, these real-world datasets are resource-intensive to create; for instance, the ApolloScape dataset required capturing over 100,000 images along with corresponding LiDAR scans and trajectories from multiple cities [7], highlighting the

practical limitations and high cost associated with real-world data annotation.

2.2 Synthetic Datasets and Simulation for Autonomy

Synthetic datasets have emerged as practical alternatives, overcoming challenges related to real-world data collection and annotation. Early synthetic datasets such as "Playing for Data" [9] and SYNTHIA [10] demonstrated the feasibility and potential of virtual environments for generating large-scale annotated data. "Playing for Data," derived from the GTA-V video game, provided 25,000 annotated frames for tasks like semantic segmentation. SYNTHIA leveraged Unity to generate a substantial virtual city dataset consisting of 213,000 images, proving synthetic datasets can effectively complement real data for training semantic segmentation models.

The Virtual KITTI datasets [5, 2] further validated synthetic approaches by carefully replicating real-world KITTI sequences in simulation, including enhanced photorealism and varied environmental conditions. Similarly, Synscapes [12] presented a photorealistic synthetic dataset of 25,000 images, highlighting the potential for synthetic data to closely approximate real-world scenarios.

CARLA, an open-source simulator based on Unreal Engine, has recently gained prominence due to its flexibility and realistic rendering capabilities. Notably, the IDDA dataset [1] leveraged CARLA to produce over a million images across multiple domains, incorporating varied weather conditions, camera settings, and diverse town layouts. This dataset is particularly used to study and benchmark domain adaptation techniques, demonstrating synthetic data's capability to mitigate domain gaps between simulation and real-world data.

2.3 Our Contribution

Our dataset advances this field by specifically addressing the scarcity of region-specific and scenario-specific annotated data through synthetic means. Utilizing the latest CARLA UE5 environment and the Upgraded Town 10 map, our dataset uniquely offers high-resolution RGB images and semantic segmentation masks formatted for YOLOv11, targeting a set of carefully selected urban driving classes. Unlike existing datasets, our work emphasizes a targeted synthetic data generation approach suitable for scenarios where obtaining real-world annotations is particularly challenging or costly. To the best of our knowledge, this is the first publicly available dataset utilizing the latest CARLA UE5 to provide semantic segmentation data explicitly tailored for urban driving scenarios at high resolution, effectively bridging the gap between synthetic training data and real-world applicability.

3 Methodology

3.1 Simulation Environment & Scenarios

Our dataset was created using the CARLA simulator, specifically the latest CARLA UE5 environment. Due to the ongoing development of the CARLA UE5 version at the time of dataset creation, customization of weather parameters was not available. Thus, all captured scenarios are under sunny, daytime conditions, providing consistent lighting conditions beneficial for initial baseline experiments and ensuring clarity in annotations.

The dataset was generated using the Upgraded Town 10 map, currently the primary environment available in CARLA UE5. Town 10 offers detailed urban environments, including complex road networks, intersections, roundabouts, and realistic urban infrastructure, ideal for semantic segmentation tasks in urban driving scenarios.

Vehicles and pedestrians were dynamically spawned within the simulation to create realistic traffic conditions. Vehicle spawning was executed using an autopilot system for diverse and realistic vehicle behavior. Specifically, 30 vehicles and 10 pedestrians were introduced into each simulation scenario, with scripts ensuring moderate density traffic and pedestrian behavior representative of typical urban conditions.

3.2 Data Generation & Annotation Process

The dataset comprises 7,233 images for training and 1,448 images for validation, captured at a resolution of 1280×720 pixels in RGB format. Each image is accompanied by semantic segmentation masks formatted specifically for YOLOv11.

Sensor data was captured using a virtual RGB

camera and a semantic segmentation camera attached to a simulated vehicle (a Dodge Charger) in CARLA. The RGB and segmentation cameras were placed at standardized coordinates relative to the vehicle (0.5 meters forward, 0 meters lateral, and 1.5 meters above the ground), offering a realistic driver-view perspective.

Data capture occurred synchronously at a fixed frame rate of 10 frames per second. Images were saved systematically, with segmentation masks being color-coded according to standard YOLOv11 classes, facilitating straightforward integration into deep learning workflows.

Figure 1 provides examples from our dataset, clearly illustrating an RGB frame captured from CARLA UE5 alongside its corresponding segmentation mask, highlighting annotated semantic classes such as cars, motorcycles, bicycles, traffic lights, buses, pedestrians, roads, sidewalks, and traffic signs. This visual representation underscores the precision and consistency of the annotations generated directly from CARLA, showcasing the practicality of synthetic datasets for tasks where high-quality real-world annotations are difficult to obtain.

4 Dataset Description and Analysis

The dataset introduced in this paper comprises a total of 8,680 images, specifically partitioned into 7,232 images for training and 1,448 images for validation. Each image is captured at a high-definition resolution of 1280×720 pixels, accompanied by precise semantic segmentation masks formatted according to YOLOv11 standards (image masks will also be included in the future with more segmented classes). These segmentation masks provide pixel-level annotations, crucial for tasks involving detailed urban scene understanding and autonomous driving perception.

4.1 Quality and Realism

The dataset leverages CARLA UE5's state-of-the-art graphical fidelity, specifically harnessing Unreal Engine 5's advanced technologies such as Lumen global illumination for realistic lighting conditions and Nanite virtualized geometry for highly detailed scene representation. Although specific post-processing effects such as motion blur were not explicitly activated, the default graphical capabilities of CARLA UE5 inherently ensure exceptional visual realism, closely approximating real-world conditions.

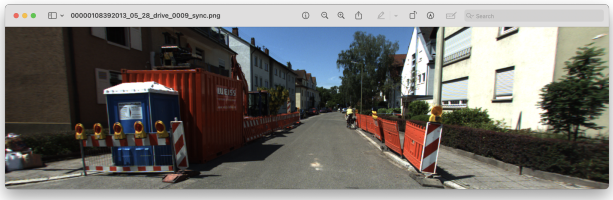

Figure 1: Example KITTI dataset image - 00000108392013_05_28_drive_0009_sync.png

Figure 2: Example dataset image label - strasbourg_000001_033027.png

Figure 3: Example dataset image label - 005695.png

For qualitative assessment, Figure 1, 2 and 3 provide visual comparisons between our dataset's synthetic images and those from established real-world benchmarks such as Cityscapes and KITTI.

This side-by-side comparison highlights the visual fidelity and practical utility of synthetic datasets produced using contemporary simulation platforms, underscoring the narrowing gap between synthetic and real data.

4.2 Class Distribution

Our dataset includes annotations for nine semantic classes considered essential for urban driving scenarios. A detailed breakdown of the class distribution in the training subset is presented in Table 1, summarizing both the frequency of occurrence and the average pixel coverage per class:

Table 1: Class distribution

Class	Count	Percentage (%)
Pedestrian	112 270	36.94
Car	79 026	26.00
Traffic Sign	62 018	20.41
Traffic Light	20 403	6.71
Bicycle	11 134	3.66
Sidewalks	7 228	2.38
Roads	7 228	2.38
Motorcycle	3 756	1.24
Bus	861	0.28

Table 1 clearly illustrates significant class imbalances, notably the underrepresentation of larger dynamic objects such as buses or motorcycles. Conversely, frequent dynamic elements like persons and cars appear extensively, yet typically cover small pixel areas due to their relatively minor visual footprint within individual images. Static elements such as roads and sidewalks consistently dominate the pixel coverage, which aligns well with typical urban scenes captured from an ego-vehicle perspective.

4.3 Sample Diversity

All dataset images were randomly selected from a substantially larger, continuously captured dataset encompassing over 150,000 frames from a single extended driving scenario spanning approximately 26 hours. The lengthy and continuous nature of this scenario ensures substantial variability in urban landscapes, traffic density, and scene complexity. Despite originating from one scenario, the prolonged duration effectively captures diverse urban driving situations, ranging from quiet streets to densely populated intersections, enhancing the dataset's representativeness and robustness for real-world applications.

4.4 Examples

Figure 4: Example dataset image - 001092.png

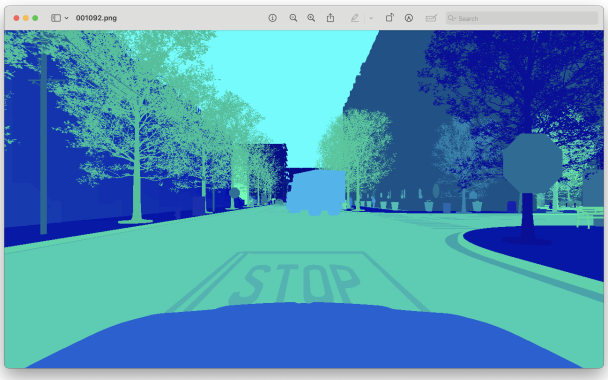

Figure 5: Example dataset image label - 001092.png

Figure 4 and 5 offer a representative example from our dataset, clearly displaying an RGB image paired with it respective semantic segmentation mask. These visual example showcase the annotation accuracy and consistency achieved through CARLA's automated labeling system, confirming the dataset's suitability for semantic segmentation tasks in autonomous driving research.

4.5 Comparison with Other Datasets

Our dataset, while smaller compared to large-scale synthetic datasets like SYNTHIA (213,000 images) or IDDA (over 1 million images), offers distinct

advantages. It emphasizes higher graphical realism provided by the latest version of CARLA UE5 and specifically focuses on a curated set of semantic classes most relevant to urban driving. Unlike Virtual KITTI, which replicates precise real-world sequences, our CARLA-based dataset introduces completely novel urban scenarios, making it uniquely complementary to existing real-world datasets and beneficial for addressing domain gaps.

4.6 Availability

The dataset is publicly available for research and academic purposes via a dedicated repository: Github Repository. It is distributed freely under an open license, with appropriate attribution required to acknowledge the original authors.

5 Discussion and Future Work

5.1 Limitations

While our dataset demonstrates significant potential for supporting urban driving perception research, it does exhibit certain inherent limitations. Primarily, despite the visual realism provided by Unreal Engine 5's Lumen and Nanite technologies, synthetic datasets, including ours, still fall short in replicating certain real-world nuances. For instance, our dataset currently lacks realistic sensor noise and detailed lens effects that are characteristic of real camera sensors, potentially limiting direct real-world applicability. Additionally, the architectural diversity is confined to the current CARLA UE5 environment, meaning certain urban environments such as highly dense city centers or rural scenarios are not represented. The absence of explicit weather variation and limited vehicle variety compared to real-world scenarios are also noteworthy constraints.

5.2 Use cases

This dataset can significantly contribute to the autonomous driving research community in various ways beyond initial semantic segmentation tasks. Researchers can leverage it for pre-training deep learning models, significantly reducing training time and resource requirements when transitioning to real-world datasets. It is particularly valuable for studies in domain adaptation, providing a robust baseline for training synthetic-to-real adaptation techniques. Additionally, due to its rich annotations, it can be utilized effectively for developing and evaluating algorithms focusing on specific urban objects such as traffic lights, signs, and pedestrian detection, potentially improving the safety and effectiveness of autonomous navigation systems in urban areas.

5.3 Future work

Future expansions of this dataset will aim to address current limitations, enhancing its breadth and applicability. Planned improvements include incorporating additional CARLA environments as they become available, introducing variations in weather and lighting conditions once these features are supported in CARLA UE5, and integrating sensor noise and lens effects to improve realism further. Additionally, future versions may include instance segmentation and depth maps, broadening the dataset's utility. We also plan to explore integrating our dataset generation approach directly into CARLA's scenario runner, enabling users to dynamically generate and tailor datasets to specific research needs without extensive downloads. We actively encourage community feedback and contributions to further enrich the dataset's applicability and effectiveness for diverse research purposes.

6 Conclusion

In summary, this paper introduces CARLA-Seg, a synthetic driving dataset comprising 8,680 high-definition images annotated with precise semantic segmentation masks using the latest CARLA UE5 simulator. By leveraging advanced graphical capabilities such as Unreal Engine 5's Lumen global illumination and Nanite virtualized geometry, our dataset effectively addresses the critical challenge of acquiring extensive and cost-intensive annotated real-world data. This synthetic dataset offers controlled yet diverse scenarios, covering essential urban driving semantic classes and facilitating robust algorithm development and testing.

The CARLA-Seg dataset enables researchers to significantly advance semantic segmentation and domain adaptation methods by providing a realistic simulation-based benchmark. It is particularly beneficial for training and evaluating models intended

for urban scene understanding, allowing algorithms to be rigorously tested under conditions that closely approximate real-world urban environments, while also supporting pre-training and fine-tuning strategies.

We publicly release the dataset to encourage its broad utilization and to stimulate further advancements within the research community. The dataset is accessible through a dedicated repository: Github Repository, and it is available under an open license requiring appropriate attribution.

Looking forward, we envision continuously improving the dataset by incorporating additional simulation features, increased environmental diversity, and supplementary data types such as depth maps and instance segmentation annotations. As simulation technology continues to mature, we anticipate synthetic datasets like CARLA-Seg becoming integral tools for developing safer and more reliable autonomous driving systems.

7 Acknowledgments

We gratefully acknowledge Aron Samaniego's support in preparing and compiling multiple simulator builds.

References

[1] L. Alberti, D. De Gregorio, and P. Soda. Idda: A large-scale multi-domain dataset for domain adaptation in semantic segmentation, 2020. arXiv preprint arXiv:2001.02279.

[2] Y. Cabon, N. Murray, and A. Gaidon. Virtual kitti 2. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 815–830, 2020.

[3] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631, 2020.

[4] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3213–3223, 2016.

[5] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4340–4349, 2016.

[6] A. Geiger, P. Lenz, and R. Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 32(11):1231–1237, 2013.

[7] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang. The apolloscape dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 954–960, 2019.

[8] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4990–4999, 2017.

[9] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 102–118, 2016.

[10] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3234–3243, 2016.

[11] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, Y. Zhang, J. Shlens, and D. Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2446–2454, 2020.

[12] M. Wrenninge and J. Unger. Synscapes: A photorealistic synthetic dataset for street scene parsing, 2018. arXiv preprint arXiv:1810.08705.

[13] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2636–2645, 2020.