Towards Using Synthetic Data in Aerial Image Segmentation

Alaa Eddine Ben Zekri, Aymen Latrach, Reza Bahmanyar
Remote Sensing Technology Institute
German Aerospace Center (DLR)
Wessling, Germany
{alaa.benzekri; aymen.latrach; reza.bahmanyar}@dlr.de

Houda Chaabouni-Chouayakh
Sm@rts laboratory
Digital Research Center of Sfax
Sfax, Tunisia
houda.chaabouni@crns.rnrt.tn

***Abstract*—This paper explores the use of synthetic datasets to improve aerial image segmentation, addressing the need for large and diverse data for model training. Current benchmarks often lack real-world conditions, such as high-altitude and nadir perspectives. To overcome this, we propose a controlled data generation approach using the CARLA simulator to generate aerial images of different towns under different weather and time of day conditions, with dynamic traffic elements. We compare our dataset with existing real and synthetic datasets, and evaluate model performance by training the DeepLabV3+ neural network on our dataset and testing on real data. The results show that incorporating synthetic data yields performance comparable to training on real data alone, highlighting its complementary value.**

***Index Terms*—Aerial Imagery, Semantic Segmentation, Synthetic Data, Deep Neural Networks, CARLA**

I. Introduction

Aerial imagery offers a unique perspective for applications such as urban planning, resource management, and social justice. These applications usually rely on in-depth analysis of aerial images using semantic segmentation as a key tool for informed decision making. Currently, semantic segmentation relies on deep learning methods [1], which have shown great promise but rely heavily on large annotated datasets that are costly and time-consuming to create, especially for complex urban environments. To address this challenge, simulators like CARLA [2] provide a cost-effective alternative, generating synthetic datasets with diverse variations in weather, lighting, and traffic conditions. Synthetic datasets have been widely used in computer vision and robotics, where domain adaptation helps models trained on synthetic data adapt to real-world images. In the aerial domain, while some synthetic datasets address bird's-eye perspectives [3, 4], to our knowledge, none have focused on replicating real-world aerial images with similar coverage, acquisition altitude, viewing angle, ground resolution, and scene complexity.

To address this gap, we use CARLA to generate a synthetic annotated aerial image dataset for semantic segmentation. To closely reflect real-world conditions and ensure image diversity, we propose a controlled data generation strategy. Our approach utilizes eight pre-built cities in CARLA, representing urban and rural environments. We capture images under eight different weather and time of day conditions, ensuring consistent variation and capturing each scene under different environmental conditions. Our dataset consists of 2.8k images, each with a size of 1408×1056 pixels and annotated with 28 object classes. Figure 1 illustrates sample images with different conditions and annotations from our dataset.

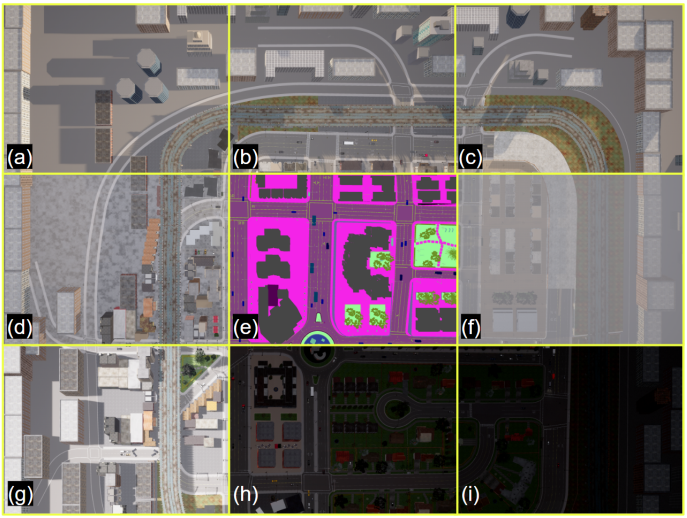

Fig. 1. Aerial image of one of CARLA Simulator's towns split into a 9×9 grid, with each patch representing a specific combination of time of day, weather conditions, and sun direction, as well as a semantic segmentation annotation. The patches are labeled from (a) to (i), with the first row (a, b, c) showing variations in sun direction under the same morning and clear weather conditions. The second row includes rainy weather (d), a semantic segmentation annotation (e), and foggy condition (f). The third row represents different times of day: Noon (g), Sunset (h), and Night (i).

We evaluate the real-world applicability of our synthetic dataset by combining it with the SkyScapes dataset [5], a real-world aerial image segmentation dataset, to train the widely used DeepLabV3+ [1] deep neural network. We then test the performance of the model on real aerial image segmentation. The results show that replacing part of the real training data with our synthetic dataset allows the model to perform comparably to a model trained solely on real data, highlighting the potential of synthetic datasets to improve model training when limited real data is available.

II. Data Generation

For our dataset, we use eight pre-built towns in CARLA, capturing images under eight different weather and time of day conditions. To simulate traffic scenarios, we populate the scenes with various vehicles, including two-wheelers, cars, trucks, and buses. To simulate the aerial perspective, we position a downward facing camera (pitched at 90°) at an altitude of 500 meters, paired with a semantic segmentation sensor covering the same field of view. The camera starts at the northwest corner of the map and moves sequentially east and south to capture non-overlapping 1408×1056 pixel images covering an area of 112 m $\times$ 84 m with a ground sampling distance (GSD) of 8 cm/pixel. Algorithm 1 shows the image generation process.

Algorithm 1 Aerial Imagery Dataset Generation

Input: H, W, XCoverage, YCoverage, FOV
Initialize: MotionBlur ← Off
for town **in** towns **do**
  for weather **in** weathers **do**
    CityBorders ← RenderTown(town, weather)
    for x **in** range(0, CityBorderEast - CityBorderWest, XCoverage) **do**
      for y **in** range(0, CityBorderNorth - CityBorderSouth, YCoverage) **do**
        vehicles ← SpawnVehiclesRandomly()
        x_position ← CityBorderWest + x
        y_position ← CityBorderSouth + y
        sensors ← PlaceSensors(x_position, y_position, H, W, FOV)
        CaptureData(sensors)
        Destroy(vehicles, sensors)
      end for
    end for
  end for
end for

III. Data Overview

Our dataset includes fine-grained semantic annotations for 28 classes as defined in CARLA's documentation [2]. For comparability with other datasets, we group these classes into seven major classes: "building", "road", "sidewalk", "vegetation", "ground", "water", and "others". Figure 2 illustrates the class distribution across towns. Urban towns, such as towns 1, 2, 3, and 10, have a high prevalence of buildings, while the other towns, with significant greenery, reflect suburban or rural characteristics. This variation demonstrates the dataset's ability to capture a diverse range of urban and non-urban environments.

In Table I, we compare our dataset with the existing datasets for semantic segmentation of aerial images. Most synthetic datasets, such as SkyScenes [4], provide bird's-eye-view images from relatively low altitudes (below 100 m), simulating common UAV images that differ significantly from aerial images captured by airborne platforms like helicopters, high-altitude UAVs, and airplanes. To ensure a fair comparison with existing synthetic datasets, we only compare our dataset to the VALID dataset [3], which contains images captured at an altitude of 100 meters. Compared to VALID, our dataset offers a larger number of aerial images and a better GSD of 8 cm/pixel, making it more suitable for precise semantic segmentation tasks. Additionally, the higher capturing altitude in our dataset reduces perspective distortion, especially for tall objects like buildings, and allows for a larger coverage area in each image, making it more effective for large-scale segmentation applications.

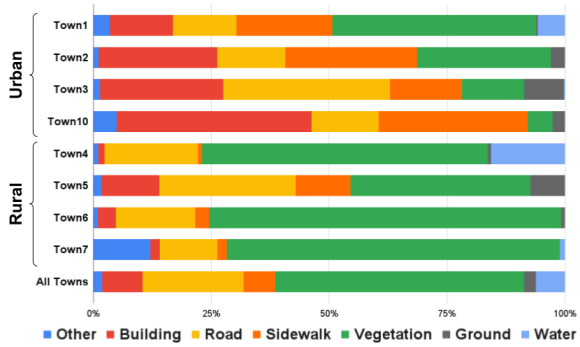

Fig. 2. Class distribution across the towns.

Compared to the real datasets, our dataset offers a higher level of annotation detail and accuracy. For example, the Aerial KITTI [7], Potsdam [6], and Vaihingen [6] datasets contain fewer classes (4 and 6 compared to our 28) and less accurate labels. Additionally, their images suffer from distortions due to imperfect orthorectification. For the TorontoCity dataset [8], the use of automated labeling across a large area compromises the precision of the labels.

In terms of scene diversity, real aerial datasets are limited by weather, daylight, and environmental conditions, as flight campaigns are typically conducted under favorable weather and specific lighting conditions. In addition, their images often lack scene diversity due to high costs and logistical challenges. In contrast, synthetic datasets such as ours and VALID allow for easy variation by using multiple virtual towns and capturing images under different weather and daytime conditions with minimal effort. Among the real datasets, only LoveDA [9] and FLAIR [10] include images from multiple towns.

IV. Experiments

We divide our dataset into training, validation, and test sets, allocating 80%, 10%, and 10% of the data, respectively, in a pseudorandom manner. To address the class imbalance between urban and rural classes, we ensure that each split contains an equal number of images from each. Additionally, we place all variations of a given image under different weather or lighting conditions in the same split. This approach results in 528 training images and 64 images each for validation and testing. By balancing urban and rural representation, we mitigate disparities in class distribution, making the dataset well suited for training and evaluating semantic segmentation models.

For the experiments, we address GPU memory limitations by splitting each image into patches of 512×512 with a 50% overlap. Additionally, we apply horizontal and vertical flipping for data augmentation.

A key measure of the quality of a synthetic dataset is the ability of models trained or tuned on it to perform

TABLE I
COMPARISON BETWEEN OUR GENERATED DATASET AND EXISTING REAL-WORLD AND SYNTHETIC AERIAL SEGMENTATION DATASETS.

Dataset	Diversity			Classes	Images	GSD (cm/pixel)	Image dimension (px)	Aerial coverage (km^2)	Altitude (m)
Diversity	**Town**	**Daytime**	**Weather**						
Real									
SkyScapes [5]	×	×	×	31	16	13	5616 × 3744	5.69	1000
Potsdam [6]	×	×	×	6	38	5	6000 × 6000	3.42	-
Vaihingen [6]	×	×	×	6	33	9	2493 × 2493 (avg)	1.36	500
Aerial KITTI [7]	×	×	×	4	20	9	variable	3.23	-
TorontoCity [8]	×	×	×	10	-	10	-	712	650
LoveDA [9]	✓	×	×	7	536	30	6000 × 6000	536	-
FLAIR [10]	✓	×	×	19	77k	20	512 × 512	817	-
Synthetic									
VALID [3]	✓	✓	✓	30	1.7k	20	1024 × 1024	-	100
Our	✓	✓	✓	28	2.6k	8	1408 × 1056	4.11	500

TABLE II
CLASS MAPPING BETWEEN OUR DATASET AND SKYSCAPES

Common Classes	Our classes	SkyScapes classes
Clutter	Other - Unlabeled - Fence - Wall - Pole - Traffic Light - Traffic Sign - Sky - Static - Dynamic - Guard Rail - Pedestrian - Train	Clutter
Urban Surface	Roads - Ground - Bridge - Rail track - Sidewalks	Paved Road - Non Paved Road - Danger Area - Bike Ways - Paved Parking Place - Non Paved Parking Place - Entrance Exit - Impervious Surface - Sidewalks
Road Markings	Road line	Lane Markings
Building	Building	Building
Car	Car	Car - Van - Trailer
Truck	Truck	Truck - Long Truck
Bus	Bus	Bus
Low Vegetation	Terrain	Low vegetation
Tall Vegetation	Vegetation	Tree

effectively on real-world data. For this evaluation, we use the SkyScapes dataset, a real-world aerial image dataset tailored for centimeter-level semantic segmentation. With a GSD of 13 cm/pixel, a coverage of 5.69 km^2, and the highest number of classes among similar real-world datasets, SkyScapes serves as a highly relevant benchmark for comparison. To enable comparison, we map the labels from our dataset and SkyScapes into a unified set of 9 classes, as detailed in Table II. While mappings for classes like "Buildings" and "Vehicles" are straightforward, differences in granularity and definitions require adjustments for others. For instance, some classes are more detailed in our dataset, while others are more nuanced in SkyScapes. Classes without equivalents, such as "Water Surface" and "Bicycle", are excluded to ensure a fair and meaningful comparison.

For our semantic segmentation experiments, we use the DeepLabV3+ [1] network, known for its strong performance on benchmark datasets like Cityscapes. The model combines techniques such as atrous convolution and an encoder-decoder structure, making it well-suited for complex urban environments in aerial imagery. We consider four training scenarios: (1) training exclusively on SkyScapes, (2) training exclusively on our dataset, (3) training on a combined dataset with 25% from SkyScapes and 75% from our dataset, and (4) training on a combined dataset with an equal 50% split from each. The first two scenarios establish baseline performances for real and synthetic data, while the latter two assess the benefits of integrating synthetic data into aerial segmentation tasks. For evaluation, we used mean Intersection over Union (mIoU), frequency-weighted IoU (FreqW IoU), and pixel accuracy (PA). These metrics are commonly used in semantic segmentation, with FreqW IoU being particularly important for addressing the significant class imbalance in both datasets.

Table III shows the test results on our dataset and SkyScapes for different training scenarios. The results show a significantly worse performance when training on SkyScapes and testing on our dataset compared to testing on SkyScapes images. A similar trend is observed when training on our dataset and testing on SkyScapes. This performance gap may be due to both the domain difference between synthetic and real data and the different scene content in each dataset. When we combine equal shares of both datasets, the performance improves significantly, with the model performing similarly well on both real and synthetic test data. Interestingly, reducing the share of real data to 25% and replacing it with our synthetic data results in performance comparable to training with three times more real data when we only use SkyScapes. This indicates the potential complementary role of synthetic data, especially when access to real-world data is limited. Figure 3 shows segmentation results for two sample images from SkyScapes across different training scenarios. Notably, models trained on the combined dataset demonstrate improved robustness to shadows, likely due to the diverse shadow orientations in the synthetic data.

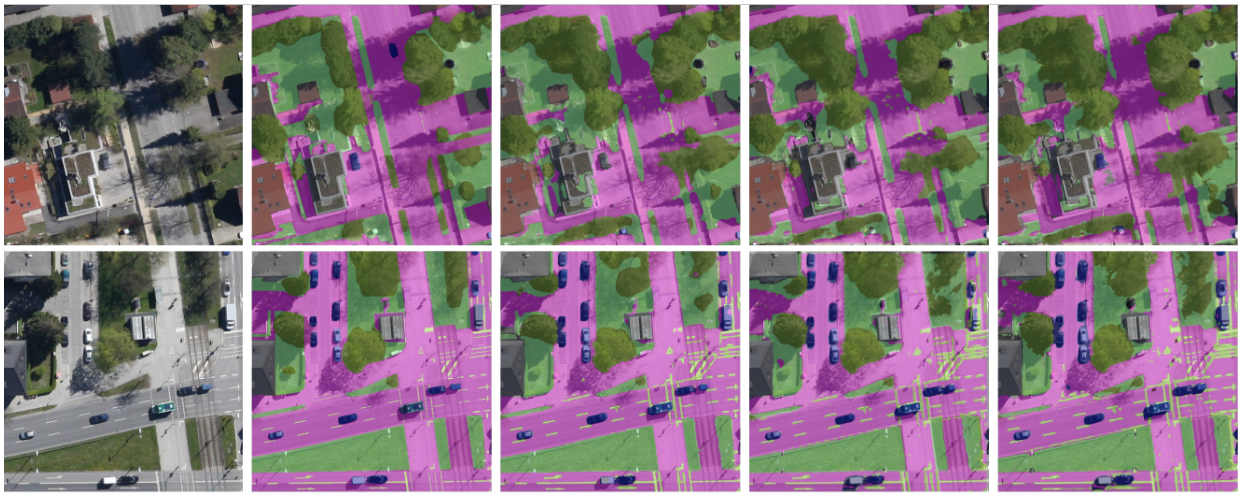

Fig. 3. Segmentation results on sample images from SkyScapes. From left to right: input images, ground truth, predictions from models trained exclusively on SkyScapes data, on the 50-50 combination of our synthetic and SkyScapes data, and on the 75% synthetic data and 25% SkyScapes data combination.

TABLE III
EXPERIMENTAL RESULTS USING DEEPLAB-V3+ ON DIFFERENT TRAINING SCENARIOS

Training	Test	mIoU (%)	FreqW IoU (%)	PA (%)
SkyScapes	SkyScapes	49.72	71.99	83.31
	Ours	17.11	34.35	43.61
50% Ours + 50% SkyScapes	SkyScapes	48.14	71.14	82.58
	Ours	58.73	66.12	73.87
75% Ours + 25% SkyScapes	SkyScapes	44.27	68.58	80.75
	Ours	53.98	73.60	79.56
Ours	SkyScapes	18.33	26.68	42.11
	Ours	57.50	75.08	82.61

V. CONCLUSION

In this work, we presented an approach for creating a diverse, large-scale synthetic dataset for aerial image segmentation using the CARLA simulator. The dataset captures high-altitude, nadir aerial perspectives with high ground resolution, and includes diverse urban and rural environments, weather conditions, and dynamic elements. Experiments with the DeepLabV3+ network show that models trained on this synthetic dataset generalize well to real-world data, especially when combined with a limited amount of real data. This highlights the potential of synthetic data to complement real data, especially when real data is scarce or difficult to obtain. Future work will focus on a more comprehensive evaluation of the dataset to further explore its potential and limitations.

ACKNOWLEDGMENT

We thank the German Academic Exchange Service (DAAD) for supporting this research.

REFERENCES

[1] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *eprint arXiv:1802.02611*, 2018.

[2] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. Ann. Conf. on Robot Learning*, 2017.

[3] L. Chen, F. Liu, Y. Zhao, W. Wang, X. Yuan, and J. Zhu, "Valid: A comprehensive virtual aerial image dataset," in *Proc. IEEE Int. Conf. Robotics and Automation*, 2020.

[4] S. Khose, A. Pal, A. Agarwal, J. Hoffman, P. Chattopadhyay *et al.*, "Skyscenes: A synthetic dataset for aerial scene understanding," *preprint arXiv:2312.06719*, 2023.

[5] S. Azimi, C. Henry, L. Sommer, A. Schumann, and E. Vig, "Skyscapes - fine-grained semantic understanding of aerial scenes," in *Proc. IEEE/CVF Int. Conf. on Computer Vision, ICCV*, 2019.

[6] ISPRS, "2d semantic labeling dataset," http://www2.isprs.org/commissions/comm3/wg4/semantic-labeling.html, [Online].

[7] G. Máttyus, S. Wang, S. Fidler, and R. Urtasun, "Enhancing road maps by parsing aerial images around the world," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2015.

[8] S. Wang, M. Bai, G. Mattyus, H. Chu, W. Luo, B. Yang, J. Liang, J. Cheverie, S. Fidler, and R. Urtasun, " TorontoCity: Seeing the World with a Million Eyes ," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017.

[9] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," in *Proc. Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

[10] A. Garioud, N. Gonthier, L. Landrieu, A. D. Wit, M. Valette, M. Poupée, S. Giordano, and B. Wattrelos, "Flair: a country-scale land cover semantic segmentation dataset from multi-source optical imagery," *eprint arXiv:2310.13336*, 2023.