SYNTHETIC DATA AUGMENTATION FOR EARTH OBSERVATION OBJECT DETECTION TASKS

Syrine Khammari
Chair of Data Processing
Technical University of Munich
syrine.khammari@tum.de

Enrique Fernández-Laguilhoat Sánchez-Biezma
FlyPix AI GmbH
enrique.fernandez@flypix.ai

Dr. Sergey Sukhanov
FlyPix AI GmbH
sergey.sukhanov@flypix.ai

Dr. Ivan Tankoyeu
AI Superior GmbH
ivan.tankoyeu@aisuperior.com

ABSTRACT

Neural networks have transformed remote sensing, making it easier to derive insights from satellite images for various Earth Observation (EO) applications. Yet, their potential is often limited by the lack of labeled data. Traditional data augmentation methods, while attempting to address this, require significant manual input and lack visual diversity, compromising model performance. We introduce an innovative data augmentation strategy that automates the generation and integration of objects into satellite imagery, enhancing datasets for object detection. Our method notably improves car detection model performance, surpassing traditional augmentation techniques.

1 INTRODUCTION

Recent improvements in satellite sensor technology and processing power, along with advancements in Computer Vision (CV), have expanded the popularity and applications of satellite imagery. Despite these advances, the practical use of object recognition and classification models in satellite imagery is often hindered by a lack of labeled data and variability, which do not fully represent the diversity found in real-world settings. This scarcity and limited representation significantly affect model performance and leave many EO challenges unaddressed.

Data augmentation (DA) has become one of the approaches to deal with the lack of sufficient amounts of labeled data Shorten & Khoshgoftaar (2019). Image-based DA includes conventional geometric transformations such as rotation, flipping, and cropping or zooming Hao et al. (2023) as well as more sophisticated methods that exploit the multi-dimensional aspect of the data Viskovic et al. (2019),Persson et al. (2018),Illarionova et al. (2021a). Some techniques exploit the capabilities of generative models Jain et al. (2021) Lin et al. (2017) Alzahem et al. (2023). Despite the performance improvement reported when using image-based DA techniques for EO, the main limitation, namely, the lack of variability remains one of the key challenges Nesteruk et al. (2022).

To address this limitation, object-based DA techniques were introduced focusing on individual objects instead of entire images and allowing more control of the synthetic images. Traditional methods in this category include copy-paste Illarionova et al. (2021b) and 3D object rendering Yan et al. (2019) techniques. The main limitation of them though is that the former generates limited visual object diversity, while the latter requires significant manual effort and expertise to create 3D models of every object often unrealistic scenes. To address object blending within the scene, Martinson et al. (2021) combines a 3D object rendering process with a domain adaptation step.

Another category of object-based DA uses Generative Adversarial Networks (GANs). Despite improved performance on various CV tasks Huang et al. (2021), these approaches remain constrained by the mode collapse problem of GANs Brock et al. (2019) and their limited generation diversity compared to likelihood models Nash et al. (2021)Child (2021).

Diffusion models Dhariwal & Nichol (2021) (DMs), a subset of likelihood models, are gaining attention for image generation in various domains with a limited focus on EO tasks e.g. in Czerkawski & Tachtatzis (2023). While not surpassing other techniques in terms of conventional performance metrics, the method yielded visually convincing results, especially for the inpainting of small areas. Another work involving DMs with EO data is the Remote Sensing Fake Sample Generation RSFSG framework Yuan et al. (2023). Conditioned on masks describing an aerial scene, their fine-tuned model was able to generate high-quality output images. Despite the quality of the generation, this approach is of less benefit for data augmentation techniques, where we want to duplicate a specific object to resolve an imbalance in the dataset. In this case, masks of the object have to be manually added to the scene in a way such as that harmony is maintained. Overall, object-based DA for EO is considered to be still a new field of research that requires attention.Current approaches either lack diversity, need a lot of manual effort or/and are limited by low image resolution.

In this work, we are filling this gap by introducing an automated object-based data augmentation method tailored to satellite imagery. Using state-of-the-art image generation models, we provide an approach that produces high-resolution labeled data with high visual diversity. Automation offers significant benefits by reducing manual effort, cutting costs, and saving time.

2 METHODOLOGY

Our approach adds synthetic objects into satellite images by inpainting masked areas, suggested by an object placement module, with objects generated by an object generation module for realism and diversity. Inputs include satellite images and text prompts for the object and its background. The proposed approach is illustrated in Figure 1. Consider a road with a car as an example. ”Road” serves as input for placement and ”Car” for generation. The result is the original image with a synthetic car on a road, assuming the former contains a road. We elaborate on both modules below.

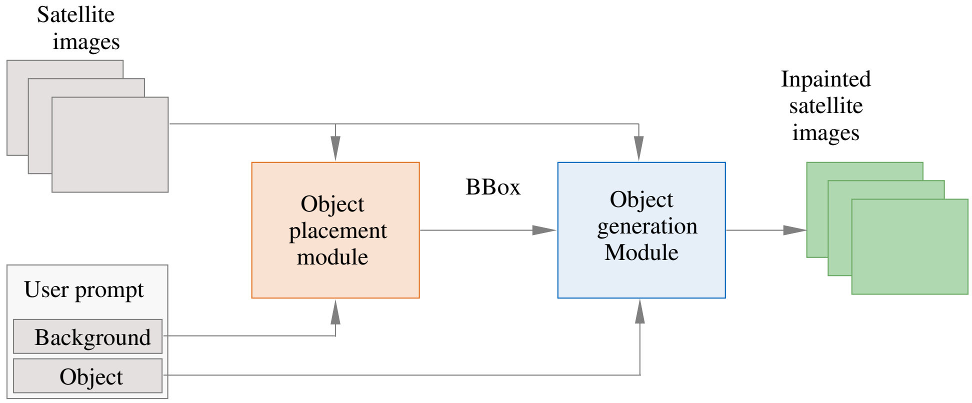


Figure 1: Overview of the data augmentation pipeline

OBJECT PLACEMENT

The object placement module determines the optimal location and size for inserting an object into an image, using a bounding box (BBox). It consists of a region proposer to identify the right area based on background type and a BBox generator for creating several non-overlapping BBoxes in that area. Users have the option to set the object size, increasing the module’s adaptability and precision in object positioning.

For the region proposer, an open-vocabulary segmentation model Zhang et al. (2023a) is used. This model takes as input a textual prompt describing the region to be segmented - a possible background on which the object can be placed (e.g. a road, a lake, etc). It outputs a mask that envelops the region. In this work, we explored two models for the segmentation of regions on satellite images. The first model, Grounded-Segment-Anything (GSAM) Ren et al. (2024), uses a two-stage segmentation approach: an open-vocabulary object detector Liu et al. (2023) to provide a visual prompt in the form of a Bounding Box (BBbox) to a promptable segmentation model called Segment Anything (SAM) Kirillov et al. (2023). The second model we consider is Segment Everything Everywhere

All at Once Model (SEEM), which is another state-of-the-art open vocabulary segmentation model that performs segmentation in a single step.

Object generation

For object generation, we use Stable Diffusion (SD) Rombach et al. (2022), a pre-trained latent diffusion model with text-conditioning. To address the domain gap between SD training data and satellite images, we fine-tuned SD with ControlNet Zhang et al. (2023b) . This enables SD to learn new generation conditioning. Specifically we chose canny edges for conditioning to ensure top views of objects.

To generate an object, we follow an inpainting method similar to the one in Czerkawski & Tachtatzis (2023). Specifically, we train ControlNet with a basic SD and combine it with an SD inpainting (SDI) model during inference. This avoids direct fine-tuning of ControlNet with SDI, which is more computationally expensive. The proposed pipeline and an inference example are shown in Figure 2.

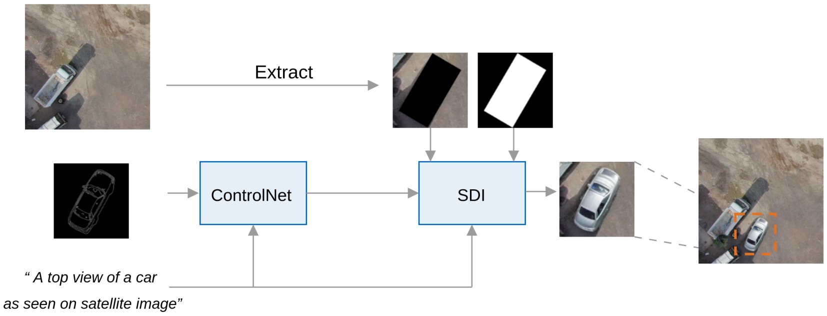


Figure 2: A pipeline for the proposed conditioned inpainting using ControlNet and SDI as well as an example of a generated object (car)

3 Evaluation and Results

To evaluate the performance of the proposed method and its components, we conducted multiple experiments presenting results for the object category cars. This object category exhibits significant visual diversity in form and color and is often required to be detected in many practical EO applications. At first, we evaluate the performance of the object generation and context-aware object placement modules. After that, we investigate how synthetically generated data is affecting the performance of a detection model. For all the experiments, we provide a link to a GitHub repository[1] with the source code and references to datasets.

Experiment 1: Object placement

In this experiment, we evaluate open vocabulary segmentation models with respect to their ability to segment specific backgrounds within satellite images. We evaluate the performance of the introduced earlier GSAM and SEEM segmentation models based on the Intersection-over-Union (IoU) metric, precision, and recall. The Domain Adaptive Semantic Segmentation (LoveDa) dataset is used due to high-resolution images for both rural and urban areas. To guide the segmentation of the LoveDa dataset, its object classes are iteratively passed to the segmentation models as text prompts e.g. *Road* or *Barren*.

Table 1: Performance metrics for GSAM and SEEM across segmentation categories

Category	**GSAM**			**SEEM**		
	IoU	**Precision**	**Recall**	**IoU**	**Precision**	**Recall**
Agriculture	0.29	0.30	0.84	0.21	0.42	0.52
Barren	0.04	0.05	0.81	0.09	0.22	0.76
Building	0.12	0.15	0.75	0.21	0.55	0.80
Forest	0.09	0.10	0.84	0.21	0.58	0.53
Road	0.15	0.21	0.62	0.19	0.76	0.51
Water	0.19	0.28	0.65	0.30	0.68	0.50

IoU metrics in Table 1 indicate both models face challenges in background segmentation, but IoU might not fully capture the region proposer's performance, given its focus on pixel precision. SEEM

[1] https://github.com/flypixai/pyxel-augment

shows strong results in precision for Road, Water, and other classes, leading to its selection for region proposal due to its conservative segmentation confirmed by recall metrics. The variations in performance can be attributed to differences in the architectures of the two models, their approaches to handling prompts, and the distinct segmentation tasks they were trained on. Our method utilizes SEEM for proposing regions but its segmentation performance across categories could restrict its applicability.

Experiment 2: Object generation

In this experiment, we evaluate the object generation quality using density and coverage metrics Naeem et al. (2020). We calculate them by analyzing synthetic and real images in feature space, using the K-nearest algorithm to find neighbours for real images. Image fidelity (density) is determined by the presence of real samples in these neighborhoods, while diversity (coverage) measures the proportion of real samples with at least one synthetic sample in their neighborhood. To fine-tune ControlNet for generating small objects such as cars, we utilized a drone dataset from Medina City (Medina-34), labeled with 34 object types, focusing on cars for this study. The dataset, publicly available, was used to fine-tune ControlNet with data points including a target image (cropped car), a source image (car's canny edges), and a text prompt describing the car's top view in satellite imagery. The training dataset contains 600 car samples. The model was trained for 34 epochs with a learning rate of 1×10^{-5}. A batch size of 2 and a gradient batch accumulation of 4 were used. Additionally, during training, 10% of the text prompts were replaced with an empty string, a common approach to encourage the model to focus on learning the visual prompt.

In our experiments, we first evaluate the quality of the fine-tuned ControlNet generation using Medina-34. In the absence of a benchmark, we compared our approach with metrics from pre-trained SD to evaluate the potential for improvement. In the appendix, we present real images, SD-generated images, and images created using our fine-tuned ControlNet. The latter consistently produces top views of cars with fewer design artefacts compared to those generated with SD. Table 2 shows that the fine-tuned ControlNet has significantly improved the quality of the generated images, both in terms of density and coverage. The table also includes metrics from additional experiments conducted to improve generation quality. One experiment improved ControlNet by using detailed text prompts for tuning and inference, while two others applied post-processing to decrease generated objects' resolution and size, notably enhancing evaluation metrics. Experiment details and prompts are available in the paper's GitHub repository.

Table 2: Density and coverage object generation: car

	Density	**Coverage**
SD	0.012	0.029
ControlNet	0.084	0.177
ControlNet with detailed prompt	0.136	0.264
ControlNet with lower resolution	0.167	0.290
ControlNet with smaller images	0.452	0.569
ControlNet (all factors combined)	**0.603**	**0.681**
Real distribution	0.993	1.0

Although the strategies in the different experiments have resulted in a significant improvement, there remains a domain gap between the synthetic and real images. To further bridge this gap, future work can explore prompt engineering for more generation guidance, for example specifying the presence of shadows, rain or snow. With the approach we have developed, we have succeeded in adding satellite data without having to manually manipulate the images. However, the process must be initiated by the user by defining the desired object and its location. It is expected that the user selects plausible locations for the objects to be placed, as our approach does not perform a plausibility check.

Experiment 3: object detection model performance with synthetic data

To assess the impact of our data generation method on object detection model performance, we fine-tune the YOLOv8 model Jocher et al. (2023), pre-trained on the COCO Lin et al. (2014) dataset. As a dataset, we are considering Pleiades Neo satellite imagery covering the Rotterdam harbour, Netherlands. Images were acquired on June 20, 2022, with a resolution of 0.29 meters/pixel. Nine object classes, including cars, were labeled, and details are accessible in our GitHub repository. For compatibility with the detection model, we used 640 × 640 crops, dividing the dataset into

”non-empty” (with cars, 20 training images with 50 objects, 211 validation images) and ”empty” (for inpainting, 20 images). We created six training sets for independent experiments: Original (20 images), Geometric augmentation ×1 (40 images), Proposed augmentation (40 images), Geometric augmentation ×2 (60 images), and Proposed + Geometric augmentation ×1 (60 images).

For every experiment, we trained on 30 epochs, set early stopping of 10 epochs, used Adam optimizer with weight decay and batch size 4, applied early stopping based on validation subset according to Mean Average Precision (mAP) metric; Several performance metrics for these experiments are averaged over three different seeds and presented in Table 3. According to Table 3 we can observe that applying geometric transformations once (total 40 images) improved the detection performance. However repeating geometric transformations on the initial dataset actually reduced performance, indicating that the model may have overfitted to the specific real instances. On the other hand the highest mAP improvement of 47% was achieved when using our proposed data augmentation approach in addition to geometric transformations. Additionally, in terms of recall and F1 score, our approach showed superior performance, achieving values of 0.275 and 0.311 respectively, surpassing the results of other approaches.

Table 3: Performance results for object detection task using YOLOv8n as backbone model

	mAP	**Recall**	**F1-score**
Original dataset	0.153	0.134	0.209
Geometric augmentation ×1	0.201	0.229	0.291
Proposed approach	0.212	**0.275**	**0.311**
Geometric augmentation ×2	0.145	0.164	0.218
Proposed approach + Geometric augmentation ×1	**0.226**	0.2158	0.293

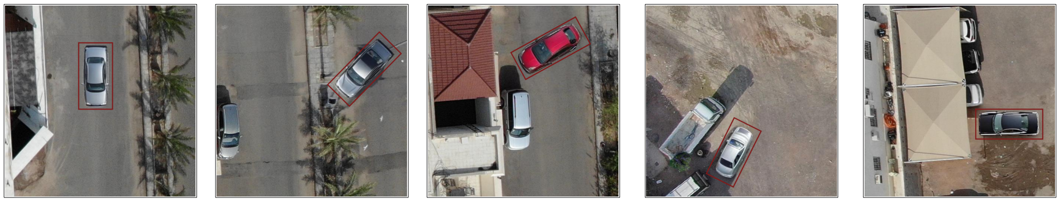


Figure 3: Examples of images generated using our proposed approach.

4 Conclusion

This paper introduces an innovative data augmentation technique for Earth Observation (EO) tasks, overcoming the issue of limited labeled data. By leveraging open vocabulary segmentation models for precise object placement and stable diffusion for generating diverse, realistic objects, our method surpasses traditional augmentation techniques in mAP and F1 scores, especially when combined with them. Future efforts will focus on refining object placement strategies for broader applicability and testing synthetic image effectiveness on various objects with sophisticated benchmarks like GANs.

Acknowledgement

The research presented in this paper was undertaken as part of Syrine Khammari’s master’s thesis. Special thanks are extended to Alice Hein for her support and guidance throughout the duration of the project.

References

Ayyub Alzahem, Wadii Boulila, Anis Koubaa, Zahid Khan, and Ibrahim Alturki. Improving satellite image classification accuracy using GAN-based data augmentation and Vision Transformers. *Earth Science Informatics*, pp. 1–18, 2023.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations*. OpenReview.net, 2019.

Rewon Child. Very deep VAEs generalize autoregressive models and can outperform them on images. In *International Conference on Learning Representations*, 2021.

Mikolaj Czerkawski and Christos Tachtatzis. Exploring the capability of text-to-image diffusion models with structural edge guidance for multi-spectral satellite image inpainting. *arXiv (arXiv:2311.03008)*, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Xuejie Hao, Lu Liu, Rongjin Yang, Lizeyan Yin, Le Zhang, and Xiuhong Li. A review of data augmentation methods of remote sensing image target recognition. *Remote Sensing*, 15(3):827, 2023.

Jian Huang, Shanhui Liu, Yutian Tang, and Xiushan Zhang. Object-level remote sensing image augmentation using U-Net-based generative adversarial networks. *Wireless Communications and Mobile Computing*, 2021:1–12, 2021.

Svetlana Illarionova, Sergey Nesteruk, Dmitrii Shadrin, Vladimir Ignatiev, Maria Pukalchik, and Ivan Oseledets. Mixchannel: Advanced augmentation for multispectral satellite images. *Remote Sensing*, 13(11):2181, 2021a.

Svetlana Illarionova, Sergey Nesteruk, Dmitrii Shadrin, Vladimir Ignatiev, Mariia Pukalchik, and Ivan Oseledets. Object-based augmentation for building semantic segmentation: Ventura and santa rosa case study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1659–1668, 2021b.

Mayank Jain, Conor Meegan, and Soumyabrata Dev. Using GANs to augment data for cloud image segmentation task. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pp. 3452–3455, 2021.

Glenn Jocher, Ayush Chaurasia, and Jing Qiu. *"YOLO by Ultralytics"*, 2023. URL `https://github.com/ultralytics/ultralytics`.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

Daoyu Lin, Kun Fu, Yang Wang, Guangluan Xu, and Xian Sun. MARTA GANs: Unsupervised representation learning for remote sensing image classification. *IEEE Geoscience and Remote Sensing Letters*, 14(11):2092–2096, 2017.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision - ECCV 2014 - 13th European Conference*, volume 8693, pp. 740–755. Springer, 2014.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv (arXiv:2303.05499)*, 2023.

Eric Martinson, Bridget Furlong, and Andy Gillies. Training rare object detection in satellite imagery with synthetic GAN images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2769–2776, 2021.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pp. 7176–7185. PMLR, 2020.

Charlie Nash, Jacob Menick, Sander Dieleman, and Peter Battaglia. Generating images with sparse representations. In *International Conference on Machine Learning*, pp. 7958–7968. PMLR, 2021.

Sergey Nesteruk, Svetlana Illarionova, Timur Akhtyamov, Dmitrii Shadrin, Andrey Somov, Mariia Pukalchik, and Ivan Oseledets. Xtremeaugment: Getting more from your data through combination of image collection and image augmentation. *IEEE Access*, 10:24010–24028, 2022.

Magnus Persson, Eva Lindberg, and Heather Reese. Tree species classification with multi-temporal sentinel-2 data. *Remote Sensing*, 10(11):1794, 2018.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Lucija Viskovic, Ivana Nizetic Kosovic, and Toni Mastelic. Crop classification using multi-spectral and multitemporal satellite imagery with machine learning. In *2019 International conference on software, telecommunications and computer networks (SoftCOM)*, pp. 1–5. IEEE, 2019.

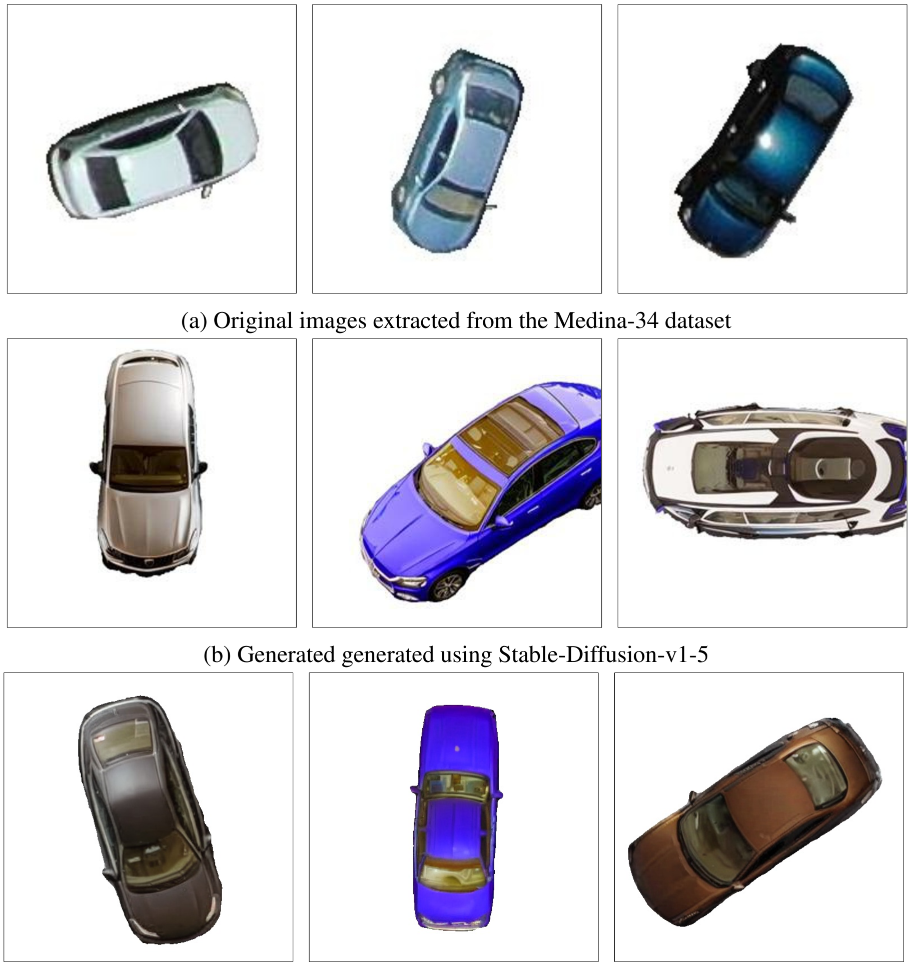
Yiming Yan, Yumo Zhang, and Nan Su. A novel data augmentation method for detection of specific aircraft in remote sensing rgb images. *IEEE Access*, 7:56051–56061, 2019.

Zhiqiang Yuan, Chongyang Hao, Ruixue Zhou, Jialiang Chen, Miao Yu, Wenkai Zhang, Hongqi Wang, and Xian Sun. Efficient and controllable remote sensing fake sample generation based on diffusion model. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1020–1031, 2023a.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023b.

APPENDIX



(a) Original images extracted from the Medina-34 dataset

(b) Generated generated using Stable-Diffusion-v1-5

(c) Generated using ControlNet fine-tuned with drone images from Medina-34 dataset

Figure 4: Example of original and synthetically generated images of cars

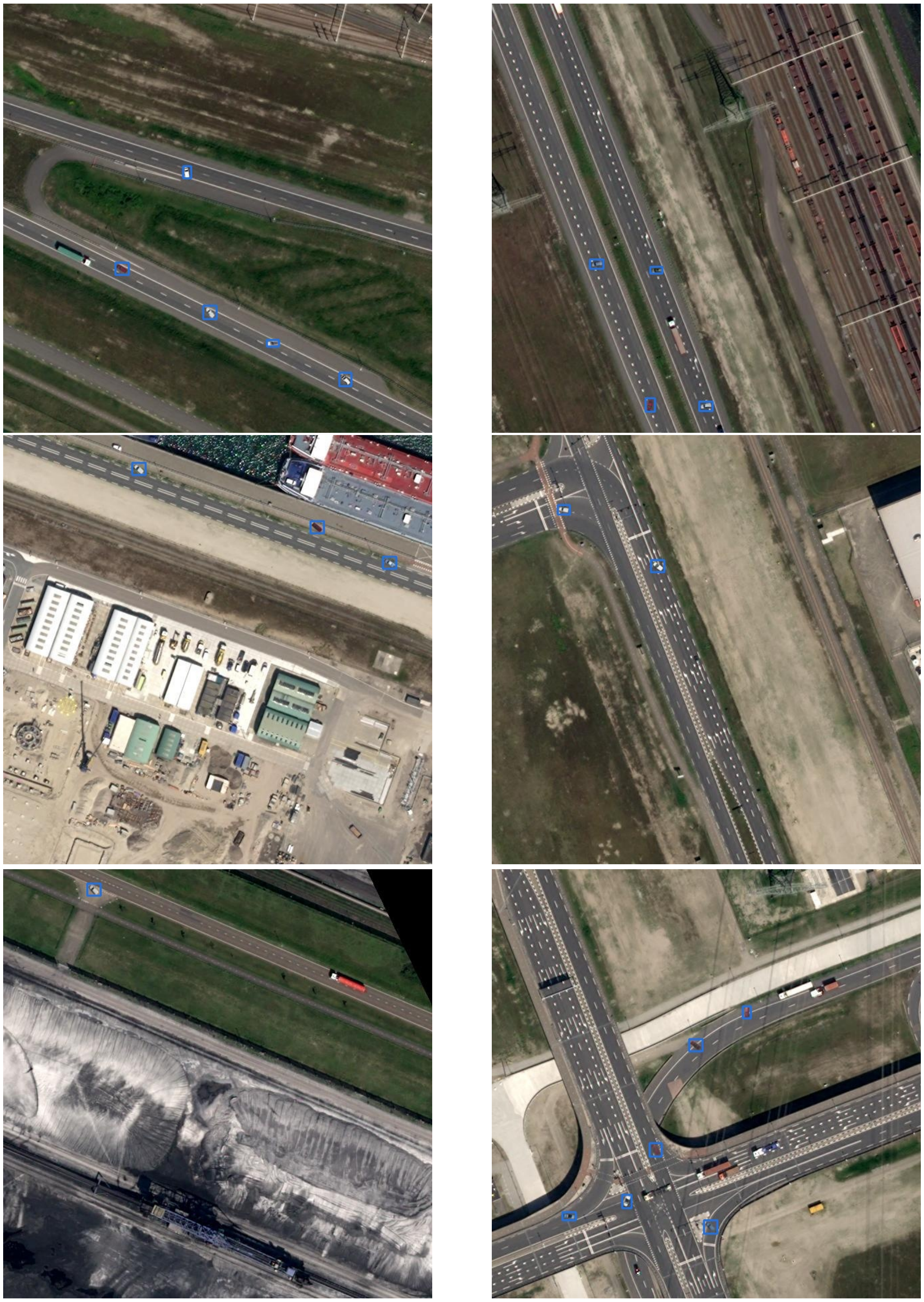


Figure 5: Subset of the synthetic dataset created using proposed object-based augmentation approach: Car on road example