Generate Your Own Scotland: Satellite Image Generation Conditioned on Maps

Miguel Espinosa
School of Engineering
University of Edinburgh
miguel.espinosa@ed.ac.uk

Elliot J. Crowley
School of Engineering
University of Edinburgh
elliot.j.crowley@ed.ac.uk

Abstract

Despite recent advancements in image generation, diffusion models still remain largely underexplored in Earth Observation. In this paper we show that state-of-the-art pretrained diffusion models can be conditioned on cartographic data to generate realistic satellite images. For this purpose, we provide two large datasets of paired maps and satellite views over the region of Mainland Scotland and the Central Belt. We train a ControlNet model and qualitatively evaluate the results, demonstrating that both image quality and map fidelity are possible. Additionally, we explore its use for the reconstruction of historical maps. Finally, we provide some insights on the opportunities and challenges of applying these models for remote sensing.

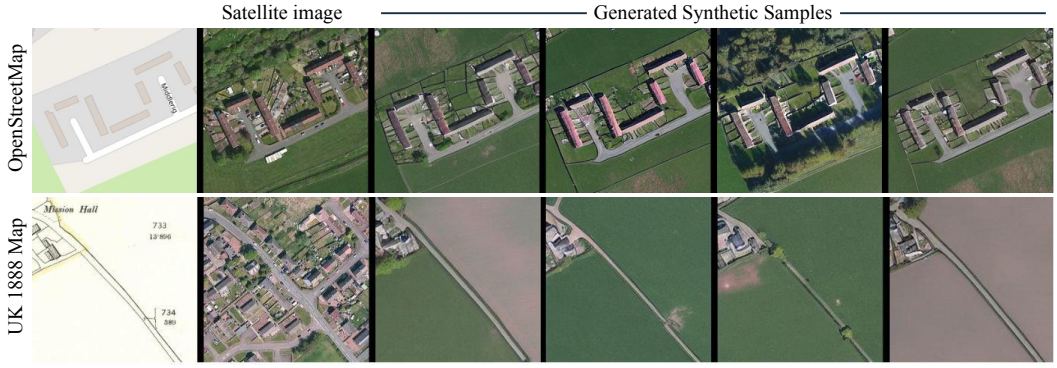

Figure 1: Examples of synthetic satellite images generated with diffusion models conditioned on OSM and UK-1888 maps (test set). The real sat. images are provided as reference (2nd column) but they are not used at inference. We cover a wide landscape diversity (urbanised and rural areas).

1 Introduction

High-resolution satellite imagery provides valuable insights into Earth's surface changes. Yet, making such images publicly available brings up privacy and legal concerns [7]. In addition, they are costly to acquire, come with increased usage restrictions for end users, and capture relatively small areas in each image. This fact, hinders the release of public datasets and slows down research in the field of Earth Observation (EO).

The generation of realistic synthetic high-resolution satellite imagery is a timely task that mitigates these concerns and opens up new possibilities. Furthermore, having a fine-grained control of the generation process allows us to create new labelled datasets without the need for human intervention for

Neural Information Processing Systems (NeurIPS) 2023 Workshop on Diffusion Models.

a wide variety of downstream tasks (object detection, segmentation, image classification, adversarial training, anomaly detection), and to augment existing ones for data-scarse situations. Such synthetic datasets enable pretraining backbone models on larger amounts of data (usually followed up by a final fine-tune stage to account for the distributional shift). In addition, recent work [14] show model improvement when training on a mixture of synthetic and real imagery.

Satellite image generation is a challenging task due to the complexity of natural scenes (fine details and textures), and the constantly varying conditions (weather, seasonality, lighting, vegetation). Only recent developments in diffusion models have enabled high-quality image generations. Moreover, when the generation is required to follow a specific layout (i.e. a map), it needs to encode its semantics, and the spatial relationship of the objects in the scene in order to synthesize a consistent image.

We generate high-quality high-fidelity satellite images conditioned on present and historical cartographic data. For this purpose we create a large dataset of paired images and train a ControlNet model. Lastly, we provide insights on the opportunities and challenges of using diffusion models for the remote sensing community.

2 Background

Generative models have significantly improved in recent years. Several works have explored their use for synthetic image generation [19], image-to-image translation [13], and data augmentation [2]. However, in the EO domain the focus has predominantly been on more traditional models, such as Generative Adversarial Networks (GANs) [9]. While GANs have shown notable results in multiple EO tasks (super resolution [8, 23], de-speckling [24], pan-sharpening [17], image generation [14, 18], haze or cloud removal [12]), they suffer from training instability and model collapse, which can lead to the generation of low quality images [5].

Diffusion models [22, 10] have emerged as promising alternatives, using stochastic processes to model the data distribution. Previous work has explored the use of diffusion models in the EO domain for diverse applications such as super resolution [16], change detection [6], and image augmentation [1]. Recent work such as ControlNet [26] allows for better control over the generation process by adding input conditions while still produce high-quality results. The use of such conditioned diffusion models in remote sensing remains unexplored, creating a gap that our work aims to address.

In the multi-modal context, previous work has explored the use of paired datasets combining different types of remote sensing data [25, 21, 11, 15]. However, the use of cartographic maps as an additional data source remains underexplored.

3 Datasets

To demonstrate the effectiveness of pretrained diffusion models in remote sensing, we construct two specific datasets for the training procedure. Instructions and code for the dataset creation can be found in `https://github.com/toastyfrosty/map-sat`. Further details are provided in Appendix A.

4 Method

We use the ControlNet [26] architecture to train a model capable of generating realistic satellite images from OSM/UK-1888 tiles. ControlNet is designed to augment pretrained image diffusion models by allowing task-specific conditioning. It has the ability to manipulate the input conditions of *neural network blocks*[1], thereby controlling the diffusion process. Intuitively, it can be seen as a way of injecting explicit guidance on the denoising process, conditioning the outputs on some reference image, in addition to the text prompt. Further details on the model architecture and on the training procedure are provided in Appedix B.

The best performing model, trained on the Central Belt dataset, is publicly available at `https://huggingface.co/toastyfrosty/controlearth`. We also publish the model trained on Mainland Scotland at `https://huggingface.co/toastyfrosty/controlearth-sct`.

[1] A *network block* in this context refers to any set of neural layers grouped as a frequently-used unit for building networks, such as a ResNet block, `conv-bn-relu` block, and transformer block, among others.

5 Analysis

We carry out a qualitative analysis of our results, mainly involving the visual inspection of the generated satellite images. This lets us evaluate more subjective elements such as colour consistency, spatial coherence and feature representation, which are often hard to quantify numerically. We include a selection of examples in Figures 1 and 2 that demonstrate the model's capabilities under different conditions (best viewed up close, in colour). More examples are provided in Appendix C.

Figure 2: Examples of synthetic sat. images generated with diffusion models conditioned on OSM and UK 1888 maps (test set). The real sat. images are provided as reference (2nd column) but they are not used at inference. Note that historical maps do not always match with the satellite images.

One of the desirable behaviours that the trained model exhibits is the diversity of samples given the same map. This shows that the model has learnt to encode the variances found in the map classes (e.g. forests), thus, successfully captured the complexity of the dataset, instead of collapsing all generations to the same image. A more detailed analysis can be found in Appendix C.

In Appendix D we compare the quality of the results when training with different datasets, and Appendix E outlines some failure cases.

5.1 Reconstruction of historical maps

As it can be seen in Figure 2, more often than not, satellite images do not match with the georeferenced historical maps due to the time difference (more than 100 years). Nonetheless, after training the model with a large "semi-paired" dataset we observe consistent generations that closely follow the map layout. These results enable the aerial and realistic reconstruction of places that could have never been captured by a satellite, being a window to past. This is of particular relevance for historians and archaeological research, aiding their visualisations and scene understanding.

6 Discussion

The use of generative diffusion models in remote sensing still remains in its early stages. However, the results presented in this study highlight their potential.

Opportunities: This approach enables the enhancement of existing datasets, by extending the number of samples. This is particularly useful for low-data regimes or scenarios where data collection can be expensive. Similarly, it can be utilised in the data augmentation step of any training pipeline. Given the diversity and realism of the generated samples it is a strong tool to ensure robustness and generalisation in models. Furthermore, the ability to synthesise high-resolution images that closely follow a specified layout (i.e. map) can be used to complement private datasets, providing a means to increase data accessibility without compromising confidentiality. Lastly, there exist multiple image-to-image use cases where this method could prove useful, for instance cloud or haze removal.

Challenges: As the quality of synthesised satellite imagery improves, concerns around misuse and the propagation of fake satellite images arise. The creation of fake satellite images or its manipulation could have harmful consequences in emergency situations, or in geopolitical events. Alongside the development of this technology, there needs to be a concurrent effort on creating regulations and ethical guidelines. On the other hand, our method is capable of creating adversarial samples (i.e. fake satellite images that resemble realistic ones), thus, it can be leveraged to create adversarial datasets. Such datasets could be used to train models for the detection of fake or manipulated satellite imagery.

Future work: The current method struggles with finer structures and undersampled classes (see Appendix E), providing room for improvement in those scenarios. Secondly, we aim to expand the current dataset by: including a wider set of modalities increasing the representational diversity (such as GIS information, DEMs, land cover data, more varied text prompts), expand its geographical coverage (to more diverse habitats and climatic regions), and develop a new sampling strategy (based on land cover maps and population density). A more complete dataset will allow for the improvement on the challenging situations across a wider range of regions. And a multi-modal dataset will enable to condition the generation process on other data modalities. Furthermore, it remains unexplored the possibilities of using different and more diverse text prompts in the generation process (for instance, for controlling seasonality changes or other weather conditions). Finally, another exciting direction is enabling consistent generation of larger maps with a smooth tiling transition. We plan to explore iterative hierarchical generation or style conditioning as possible methodologies to achieve this objective. Such method would open possibilities for artists and content creators.

7 Conclusion

We have demonstrated that state-of-the-art diffusion models can be used to generate realistic satellite images conditioned on maps. For this purpose, we create a large dataset containing pairs of maps and satellite images for Mainland Scotland and the Central Belt regions. We successfully train ControlNet models and provide insights on the results obtained. Finally, we outline some possible directions for improvements, and discuss the potential of generative methods in the field of EO.

References

[1] Oluwadara Adedeji, Peter Owoade, Opeyemi Ajayi, and Olayiwola Arowolo. Image Augmentation for Satellite Images. *arXiv*, July 2022.

[2] Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data Augmentation Generative Adversarial Networks. *arXiv*, November 2017.

[3] World Imagery (Clarity), February 2022. [Online; accessed 18. Jul. 2023].

[4] World Imagery, July 2023. [Online; accessed 18. Jul. 2023].

[5] Martin Arjovsky and Leon Bottou. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*, 2017.

[6] Wele Gedara Chaminda Bandara, Nithin Gopalakrishnan Nair, and Vishal M. Patel. DDPM-CD: Remote Sensing Change Detection using Denoising Diffusion Probabilistic Models. *arXiv*, June 2022.

[7] Megan M. Coffer. Balancing Privacy Rights and the Production of High-Quality Satellite Imagery. *Environ. Sci. Technol.*, June 2020.

[8] Bekir Z. Demiray, Muhammed Sit, and Ibrahim Demir. D-SRGAN: DEM Super-Resolution with Generative Adversarial Networks. *SN Computer Science*, 2(1):1–11, February 2021.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

[11] Danfeng Hong, Lianru Gao, Naoto Yokoya, Jing Yao, Jocelyn Chanussot, Qian Du, and Bing Zhang. More Diverse Means Better: Multimodal Deep Learning Meets Remote-Sensing Imagery Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 59(5):4340–4354, August 2020.

[12] Anna Hu, Zhong Xie, Yongyang Xu, Mingyu Xie, Liang Wu, and Qinjun Qiu. Unsupervised Haze Removal for High-Resolution Optical Remote-Sensing Images Based on Improved Generative Adversarial Networks. *Remote Sensing*, 12(24):4162, December 2020.

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[14] Van Anh Le, Varshini Reddy, Zixi Chen, Mengyuan Li, Xinran Tang, Anthony Ortiz, Simone Fobi Nsutezo, and Caleb Robinson. Mask Conditional Synthetic Satellite Imagery. *arXiv*, February 2023.

[15] Jiaxin Li, Danfeng Hong, Lianru Gao, Jing Yao, Ke Zheng, Bing Zhang, and Jocelyn Chanussot. Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, 112:102926, August 2022.

[16] Jinzhe Liu, Zhiqiang Yuan, Zhaoying Pan, Yiqun Fu, Li Liu, and Bin Lu. Diffusion Model with Detail Complement for Super-Resolution of Remote Sensing. *Remote Sensing*, 14(19):4834, September 2022.

[17] Jiayi Ma, Wei Yu, Chen Chen, Pengwei Liang, Xiaojie Guo, and Junjun Jiang. Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion. *Information Fusion*, 62:110–120, October 2020.

[18] Javier Marín and Sergio Escalera. SSSGAN: Satellite Style and Structure Generative Adversarial Networks. *Remote Sens.*, October 2021.

[19] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016.

[20] 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*, pages 10684–10695, 2022.

[21] Manish Sharma, Mayur Dhanaraj, Srivallabha Karnam, Dimitris G. Chachlakis, Raymond Ptucha, Panos P. Markopoulos, and Eli Saber. YOLOrs: Object Detection in Multimodal Remote Sensing Imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:1497–1508, November 2020.

[22] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[23] Hai Sun, Ping Wang, Yifan Chang, Li Qi, Hailei Wang, Dan Xiao, Cheng Zhong, Xuelian Wu, Wenbo Li, and Bingyu Sun. HRPGAN: A GAN-based Model to Generate High-resolution Remote Sensing Images. *IOP Conference Series: Earth and Environmental Science*, 428(1):012060, January 2020.

[24] Puyang Wang, He Zhang, and Vishal M. Patel. Generative adversarial network-based restoration of speckled SAR images. In *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 1–5. IEEE, December 2017.

[25] Xin Wu, Danfeng Hong, and Jocelyn Chanussot. Convolutional Neural Networks for Multimodal Remote Sensing Data Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–10, November 2021.

[26] Lvmin Zhang and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. *arXiv*, February 2023.

A Datasets description

The multi-modal dataset pairs 256×256 OSM/UK-1888 image tiles with corresponding 256×256 World Imagery [4] satellite image tiles. We use a fixed text prompt "*convert this openstreetmap/old map into its satellite view*" for the pretrained SD model. The area considered in this study is mainland Scotland. The sampling strategy consists of random sampling over a predefined region.

We carry out experiments on multiple datasets, that is, sampling across different regions (Figure 3): (1) all of Mainland Scotland, and (2) the Central Belt region. The motivation behind sampling across different regions is to account for unbalanced geographic features; Mainland Scotland is dominated by rural areas, forests, mountain ranges, and fields whereas the Central Belt region has a much larger representation of human-made structures like buildings, roads, and other features found in larger cities. The Mainland Scotland dataset contains 78,414 training pairs of images, and the Central Belt dataset 68,195 training pairs (with an additional 20% of test pairs for each case).

Figure 3: Sampling regions used for the dataset construction. We visualise some pair examples (map, satellite img). Mainland Scotland is largely rural, whereas the central belt has build up cities including Edinburgh and Glasgow.

We use OpenStreetMap tiles, UK-1888 map tiles and World Imagery satellite images, both at a zoom level of 17. For the central belt region, we explore two products from the free World Imagery service as provided by ArcGis Online: the latest World Imagery version and the older Clarity version (deprecated) [3]. We find that the Clarity version retains more detail and higher image quality, so we train our models on both versions for a comparative evaluation (note that World Imagery products are composites compiled from different sources and providers, resulting in varying resolutions across locations).

B ControlNet details

Given a feature map $x \in \mathbb{R}^{h \times w \times c}$ where $\{h, w, c\}$ represent height, width, and channel numbers respectively, a neural network block $\mathcal{F}(\cdot; \theta)$ with a set of parameters θ transforms x into another feature map y via the relation $y = \mathcal{F}(x; \theta)$.

Crucially, as Figure 4 illustrates, ControlNet [26] keeps the parameters θ locked, cloning it into a trainable copy θ_c which is trained with an external condition vector c. The idea behind making such copies instead of directly training the original weights is to mitigate overfitting risks in small datasets and being able to reuse larger models trained on billions of images.

An important innovation is the introduction of a *zero convolution* layer to connect the frozen network blocks and the trainable copies (Figure 4). Zero convolution is a 1×1 convolution layer with both weight and bias initialised as zeros. Note that ControlNet initially will not affect the original network

at all, but as it is trained, it will gradually start to influence the generation with the external condition vectors.

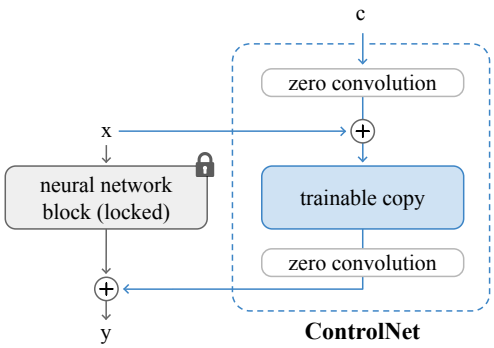

Figure 4: ControlNet network blocks with "zero convolutions" (1×1 convolution layer with both weight and bias initialised to zeros). Figure adapted from the original work [26].

B.1 ControlNet for Satellite Image Synthesis

We use the ControlNet architecture, along with a large pretrained diffusion model (Stable Diffusion) to translate OpenStreetMap images into realistic satellite images.

We follow the same training process as in the original ControlNet architecture [26]. Our model progressively denoises images in the perceptual latent space to generate samples. It learns to predict the noise added to the noisy image, and this learning objective is used in the fine-tuning process of the entire pipeline.

As the Stable Diffusion (SD) [20] weights are locked, the gradient computation on the SD model can be avoided, which accelerates the training process and saves on GPU memory. Leveraging a large pretrained diffusion model not only improves computational efficiency, but also yields higher-quality results.

B.2 Training and inference details

We carry out multiple experiments with different pretrained large diffusion backbones. Specifically we experiment with two different versions from Stable Diffusion: v.1-5, and v.2-1. We find that SD version v.1-5 tends to give better results. Experiments are run on a cluster node of 8 A100 40GB GPUs. The batch size is set to 2048 for 250 epochs. The training time is approximately 8 hours and the learning rate is kept constant at 0.00001. During inference, images are sampled with 50 inference steps (further increasing the number of inference steps doesn't have a noticeable impact on image quality), and it takes 2-3 seconds per image.

C Examples of generated satellite images

Rows 1-4 in Fig. 5 illustrate seasonality changes in the different samples. Similarly, other variances are also perceivable, such as weather phenomena, lighting conditions and human activity. Rows 5-8 are examples for water bodies of multiple sizes, such as rivers, human-made canals, and open sea in coastal regions. Lastly, rows 9-11 show urban areas and more elaborate human-made patterns which the model is able to closely follow.

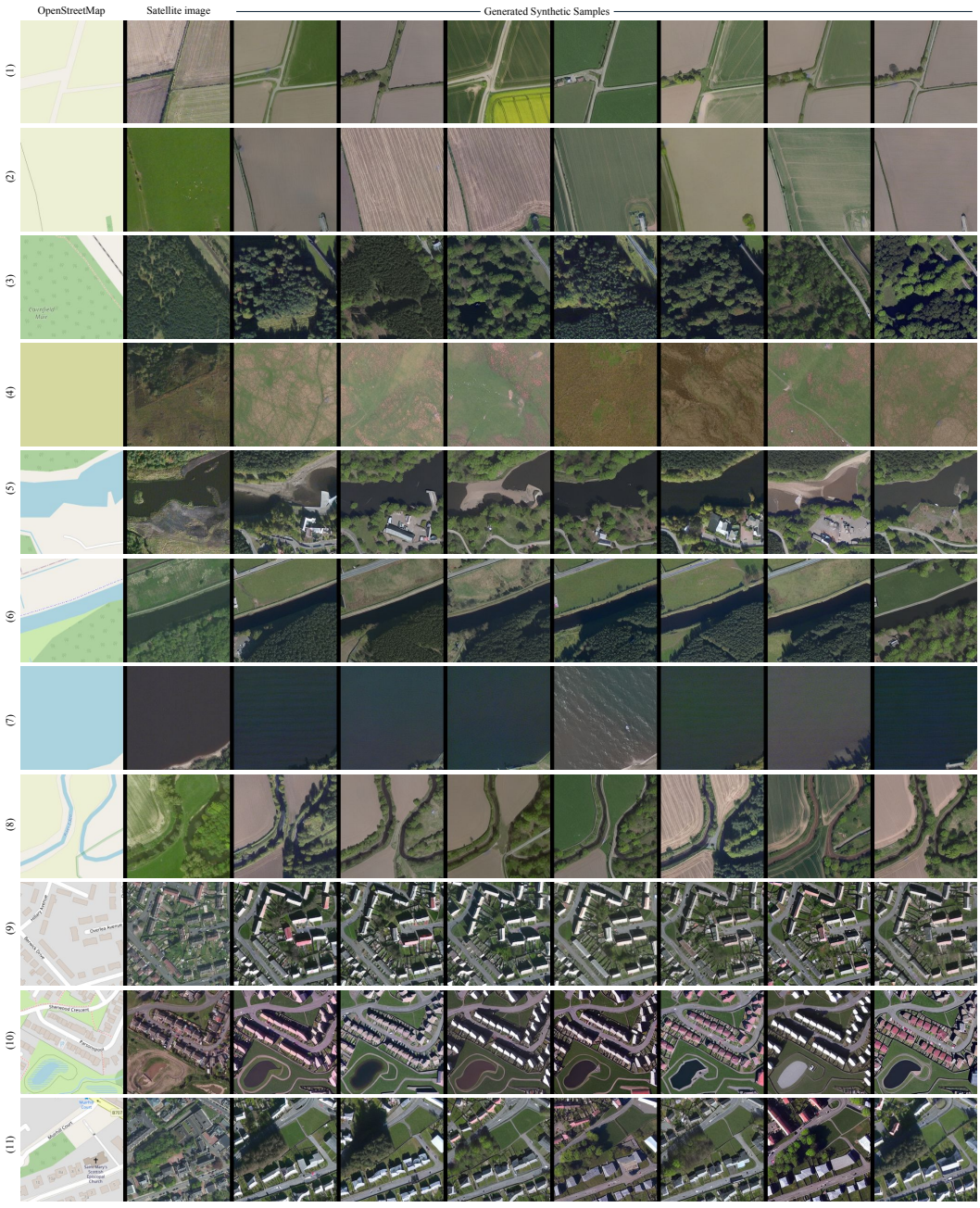

Figure 5: Examples of synthetic satellite images from the trained model conditioned on maps. All images shown correspond to the test set. The real satellite images are provided as reference (second column) but they are not used at inference. Rows 1-4 show agricultural land, forests and bare areas. Rows 5-8 illustrate water bodies at varying sizes. Rows 9-11 correspond to different man-made structures, which condition the generation with more intricate patterns.

D Dataset quality comparison

As discussed in Section 3, we train the same model on two different versions of the central belt dataset (one with World Imagery updated product, and the other with the deprecated Clarity version). Figure 6 provides a comparative visual analysis of two identical ControlNet models, both subjected to the same training parameters but on the two distinct datasets. As it can be observed, the deprecated Clarity product shows finer details and superior image quality. Therefore, it becomes evident that the quality of the learned representations is heavily influenced by the quality of the training data employed.

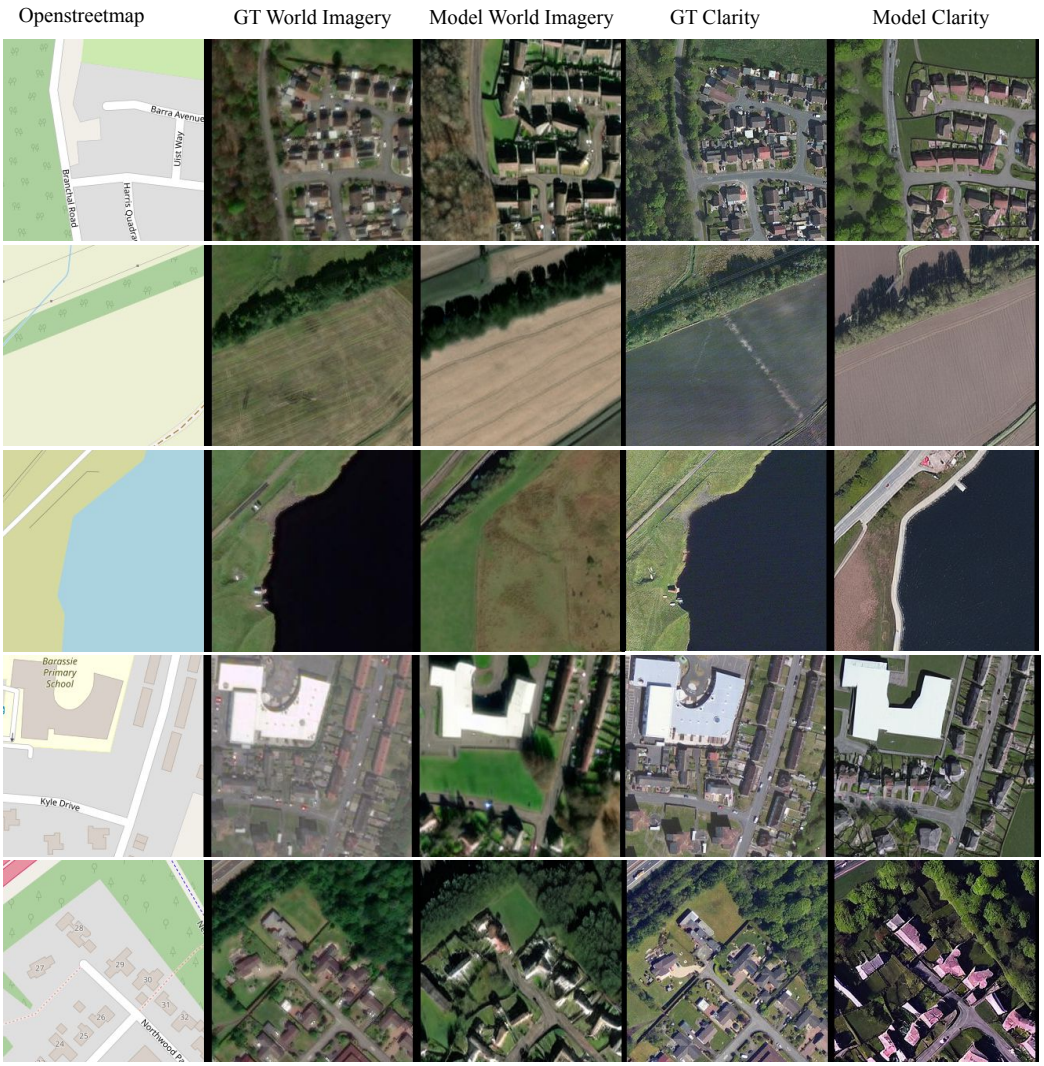

Figure 6: We illustrate the quality differences when training the same model with World Imagery and Clarity datasets over the Central Belt area. GT stands for Ground Truth, i.e. the real satellite images.

E Failure cases

Some failure cases are shown in Figure 7. Large roads, specially those with lanes and straight lines are found challenging by our model. Equally, intersections and road overpasses are difficult to generate coherently. Rivers are easily mistaken by roads in some of the samples, and we show a failure case for a larger water body, where it is confused by a building (possibly due to its polygonal shape). Lastly, we also visualise railroads as challenging scenarios. These occurrences can largely be attributed to the under-representation in our dataset of the specific scenarios.

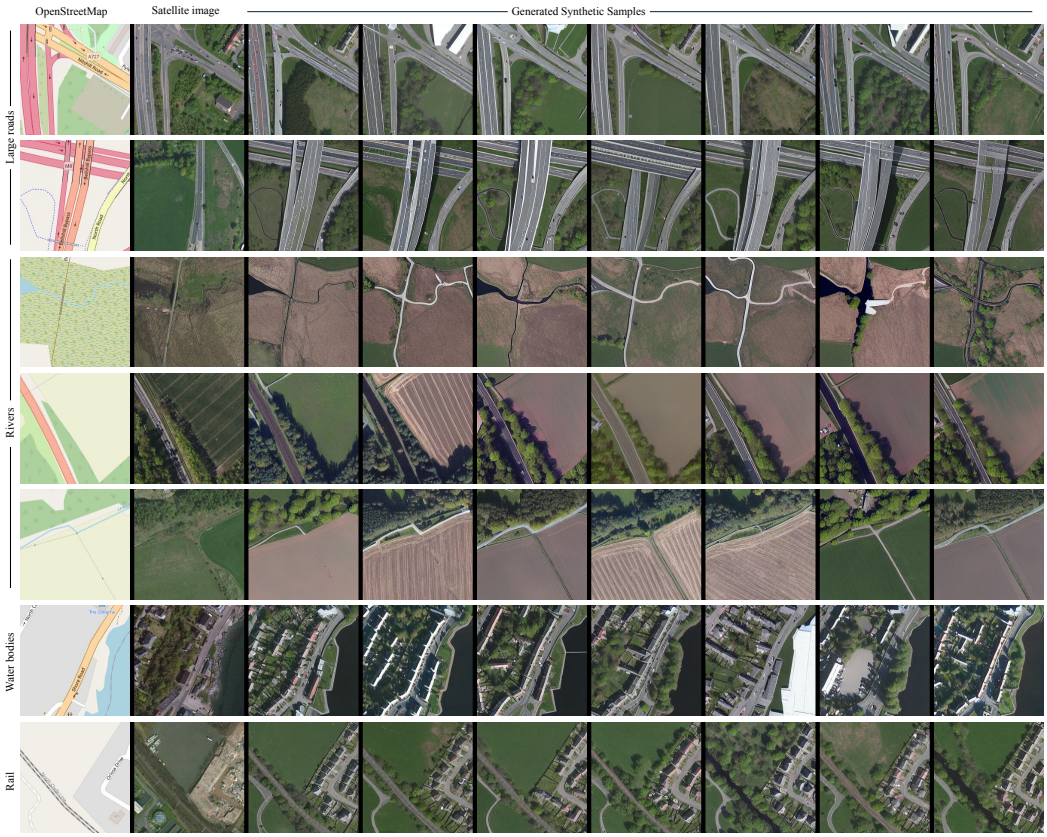

Figure 7: Failure cases for more challenging scenarios (which usually correspond to under-represented cases in the dataset, such as larger railways, coastal regions, or road intersections). The real satellite images are shown in the second column for comparative purposes.