

# Supplementary Material for CVPRW 2025 paper MORSE-015

Paul Borne--Pons<sup>1,2\*</sup> Mikolaj Czerkawski<sup>2,3</sup> Rosalie Martin<sup>1</sup> Romain Rouffet<sup>1</sup>  
<sup>1</sup>Adobe Research <sup>2</sup>European Space Agency (ESA) <sup>3</sup>Asterisk Labs

## A. Further Output Examples

More output examples are shown in Figure 1.

## B. Implementation details

Further details on the dataset generation and training are provided below.

### B.1. Dataset Details

More examples of training data are shown in Figure 2.

#### B.1.1 Pre-processing and Augmentations

We used minimal augmentations since the dense sampling of Earth terrains already acts as an augmentation method. Spatially close terrains often share the same captions but exhibit slightly different features, providing natural variability. For both modalities, we extracted 4 regular crops from each Major TOM cell and resized them to match the input width and height required by the pre-trained variational autoencoder (VAE) (768). We explicitly avoided using crops of varying sizes, as this could result in inconsistent object sizes within the generated images.

Major TOM thumbnails (B04, B03, and B02 scaled reflectances that were processed by applying an appropriate gamma curve before clipping values to the range [0, 1]) were then directly used for RGB inputs.

The DEM data was stacked three times along the channel axis to match the input channel size of the VAE. Since much of Earth’s surface is flat, directly normalizing the DEM based on global extremes (max elevation = 8,848 m and min elevation = -420 m) led to mode collapse during training (outputting flat maps for almost all captions). To address this, we normalized the DEM values relative to their local range, scaling the DEM between its own maximum and minimum elevations.

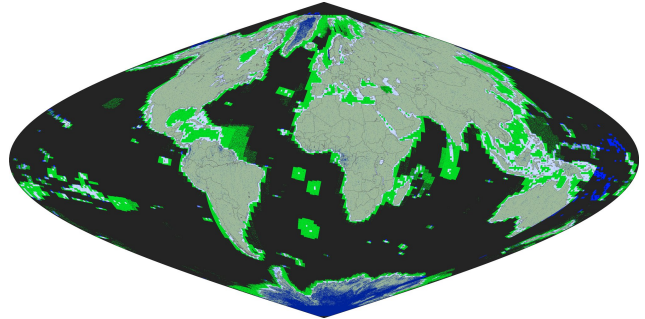


Figure 1. Coverage of the dataset used for this work. Every pixel corresponds to a single cell on the Major TOM grid (10 km). Green marks regions with only Sentinel-2 images available, while blue indicates those with only DEM. Black indicates the absence of any data, while the land and water colors represent the presence of both modalities.

#### B.1.2 Filtering the dataset

To close the gap between the list of all tiles with available RGB and DEM seen in Figure 1 and the finite dataset, one needs to exclude quite a few cells.

- Oceans and large inland water bodies, where the DEM is flat and the RGB texture-like.
- Leaving out Antarctica introduce a bit of bias in the model but heavy distortion due to the proximity to the poles in the DEM images (natively in EPSG:4236) and intense clipping of the thumbnails due to the very important reflectance of the snow that covers most of the continent make the samples greatly hinders the quality of the samples. The same effect can be seen on the Northernmost images of Canada and Russia, quite well represented in the dataset (the two biggest countries in the world with respectively 11% and 9% of the total world’s landmass and) suffering from similar distortions.
- Cells where RGB and DEM don’t have the same projection (close to the limit of the UTM zones).
- Cells where more than 0.8 percent of the image is missing, this happens near the limit of UTM zones (cropping during the creation of the MT dataset)

---

\*First author

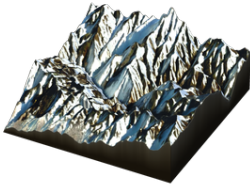
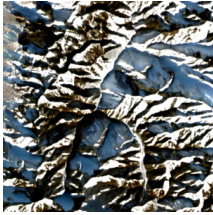
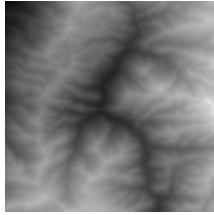
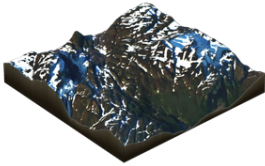
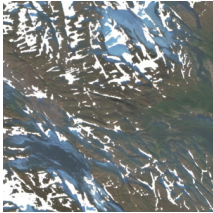
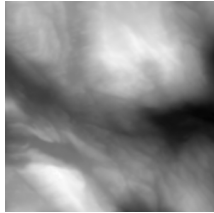
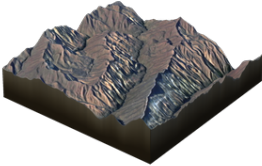

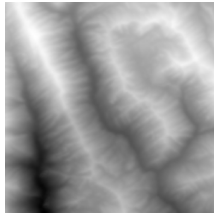
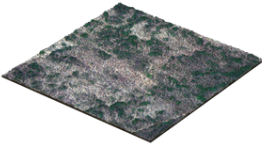
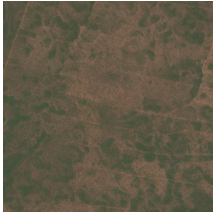
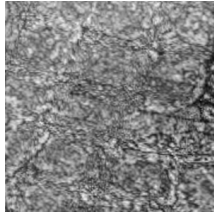
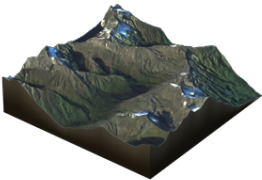
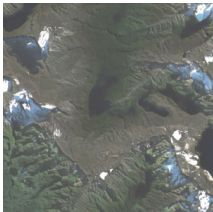
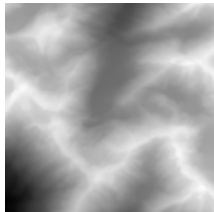
Terrain	RGB	DEM	Caption
			"mountains in the Alps in january"
			"norway fjords"
			"puna and mountains in Bolivia in May"
			"savanna and plains in Brazil in January"
			"temperate forests and mountains in New Zealand in November"

Table 1. Samples from the model, the Terrains were created using the RGB and DEM maps in Substance Designer. Each caption was prefixed by "A sentinel 2 image" before being fed to the model.

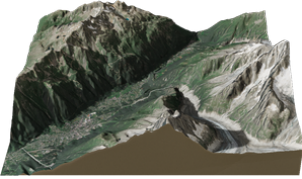
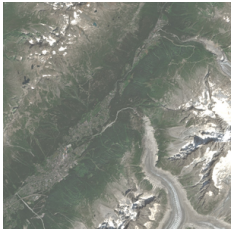
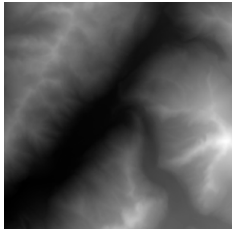
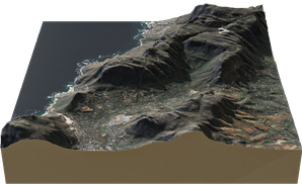

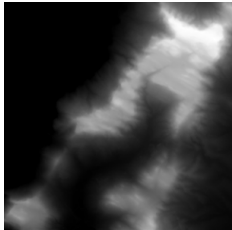
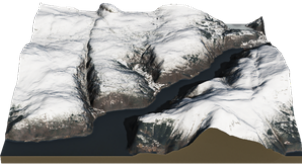
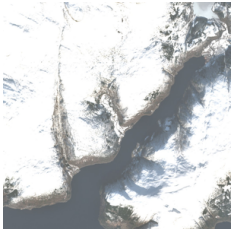
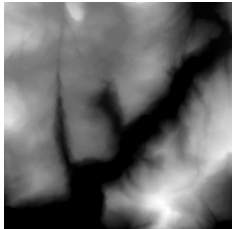
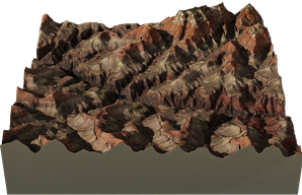
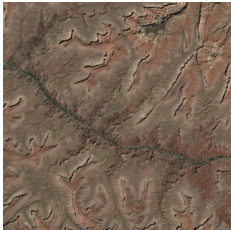
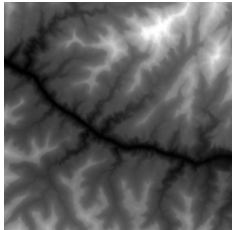
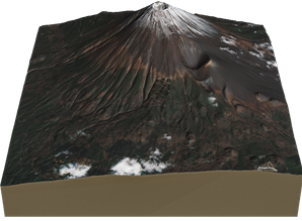
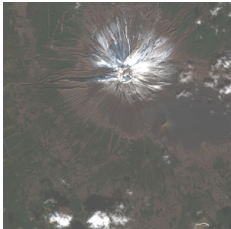


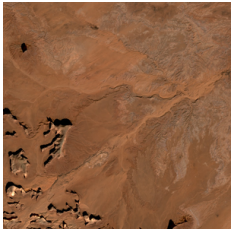
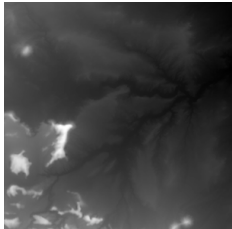
Name (Major TOM ID)	Terrain	RGB	DEM	Caption
Chamonix (511U_53R)				"conifer mixed forests and Alps in France in July"
Tabletop Mountain (379D_169R)				"shrubland and mountains in South Africa in April"
Novegian Fjord (673U_36R)				"Scandinavian Montane Birch forest and grasslands and Norway in Norway"
Grand Canyon (401U_1009L)				"shrublands and mountains in United States of America in May"
Mount Fuji (393U_1260R)				"alpine conifer forests and mountains in Japan in November"
Monuments Valley (411U_980L)				"shrublands and mountains in United States of America in February"

Table 2. Samples from the dataset, the Terrains were created using the RGB and DEM maps in Substance Designer. Each caption was prefixed by "A sentinel 2 image" before being fed to the model.



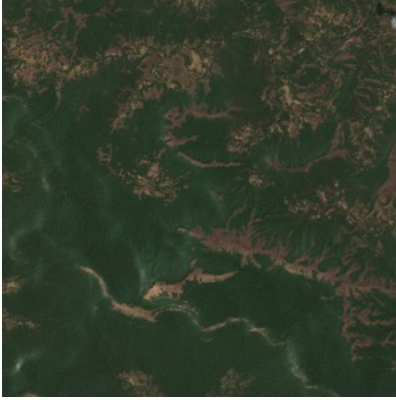


Figure 2. S2 L2A image



Figure 3. S2 L1C image

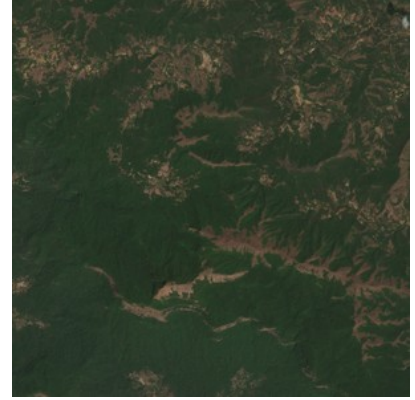


Figure 4. Corrected image

Figure 5. Histogram Matching: Example on Major TOM tile 154U 809R. The L2A product has overcorrected shadows and the L1C product has haze.

### B.1.3 Correcting shadows

Sentinel-2 satellite images are available in two processing levels:

- L1C (Top-Of-Atmosphere, TOA): These images include atmospheric effects such as scattering and absorption, making them useful for cloud detection.
- L2A (Bottom-Of-Atmosphere, BOA): These images are corrected for atmospheric effects using the Sen2Cor pipeline. This correction attempts to remove shadows and shading by assuming a Lambertian surface. However, this over-simplistic model often results in over-corrected shadows that are visually unappealing.

Shadows play a crucial role in the joint representation of RGB images and Digital Elevation Models (DEMs), as they embed spatial structure of DEMs into the RGB images. A simple yet effective way to mitigate over-corrected shadows is to apply histogram matching between the L1C and L2A products. This process "dehazes" the L1C image by transferring the color distribution of the L2A product onto it. Since histogram matching is a non-decreasing transformation, the shadows present in L1C are preserved in the corrected image.

Two models were trained:

- A first model leveraging the joint representation of Sentinel-2 L2A and Copernicus DEM.
- A second model trained on the shadow-preserving, histogram-matched version of the images.

### B.2. Model Implementation Details

We used stable-diffusion-2-1 (768 input shape trained with *v*-objective and Exponential Moving Average) from HuggingFace and used the scripts from the diffusers library as a base for our implementation. To match the base model we train with  $768 \times 768$  inputs and a *v*-objective. We use

8 NVIDIA A100 (40GB) GPUs. The text-to-terrain model was trained with a batch size of 128 for 80000 iterations, which we determined was sufficient for convergence. We choose a constant learning rate of  $1e-5$  with the AdamW optimizer. For sampling, we use the DDIM [4] sampler with 30 steps and a guidance scale of 7.

## C. Ethical Concerns & Carbon footprint

### C.1. Ethical concerns

Generative AI models for image creation raise several ethical concerns, including issues of artists' data ownership, potential misuse for creating fake media, biases in training data, and their impact on artists' livelihoods. Concerns about data ownership and bias usually stem from the use of web-crawled datasets with unclear copyright licenses, our training data is quite free of such concerns being built on global satellite images but such problems could arise from the Stable Diffusion 2.1 weights we fine-tuned on.

While these technologies may streamline creative processes, they must not replace human expertise for nuanced control and customization, leaving room for collaboration rather than outright replacement. This underlines especially the need for further work on conditioning of the models following methods similar to [1, 2] or [3].

### C.2. Carbon footprint

Aggregating all the training we used, in total, 8 A100 GPU 40GB during 200 hours. Considering the data center used for computation it amounts to about 80kg CO<sub>2</sub> (which is equivalent to a Paris-London flight).

## References

- [1] Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David Lobell, and Stefano Ermon. DiffusionSat: A Generative Foundation Model for Satellite Imagery, Dec. 2023. [4](#)
- [2] J. Lochner, J. Gain, S. Perche, A. Peytavie, E. Galin, and E. Guerin. Interactive Authoring of Terrain using Diffusion Models. *Computer Graphics Forum*, 42(7):e14941, Oct. 2023. [4](#)
- [3] Srikumar Sastry, Subash Khanal, Aayush Dhakal, and Nathan Jacobs. Geosynth: Contextually-aware high-resolution satellite image synthesis, 2024. [4](#)
- [4] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models, Oct. 2022. [4](#)