DIFFUSION MODELS FOR EARTH OBSERVATION USE-CASES: FROM CLOUD REMOVAL TO URBAN CHANGE DETECTION

Fulvio Sanguigni[1,2], *Mikolaj Czerkawski*[1], *Lorenzo Papa*[2], *Irene Amerini*[2], *B. Le Saux*[1]

[1]Φ-lab, ESRIN, European Space Agency, Frascati I-00044, Italy
[2]Department of Computer, Control and Management Engineering, AlcorLab
Sapienza University of Rome, Via Ariosto 25 00185, Rome, Italy

ABSTRACT

The advancements in the state of the art of generative Artificial Intelligence (AI) brought by diffusion models can be highly beneficial in novel contexts involving Earth observation data. After introducing this new family of generative models, this work proposes and analyses three use cases which demonstrate the potential of diffusion-based approaches for satellite image data. Namely, we tackle cloud removal and inpainting, dataset generation for change-detection tasks, and urban replanning.

***Index Terms*—** Generative modelling, Diffusion Models, Cloud Removal, Change Detection, Urban Planning, Image Inpainting

1. INTRODUCTION

After a decade of Artificial Intelligence (AI) pervading all aspects of Earth observation [1] thanks to discriminative models being able to generate products out of satellite data such as classification maps, here comes the age of generative models able to directly encode the nature of Earth observation (EO) data. After Generative Adversarial Networks [12, 2] and Energy-based models [3], the current state-of-the-art for generative models is largely achieved via denoising diffusion. The diffusion-based solutions gave rise to advancements such as text-to-image generators [13, 14], super-resolution models [15], or other image inverse tasks [16]. The high generative capability of these techniques could potentially be transferred for Earth observation tasks, however, the exact applications in this context are currently sparse.

To provide a perspective on how these models could bring benefits for Earth observation data, several use cases are introduced and demonstrated herein. The three use cases cover a wide range of domains where diffusion models, and more specifically, diffusion-based inpainting, are applicable. This includes a use case of cloud removal already explored in the literature, a delivery of a synthetic change detection dataset, and finally, a use case of urban replanning with satellite imagery. In the following we present and explain Diffusion Models in Section 2 then explore the three use cases in Section 3 before drawing a few perspectives.

2. BACKGROUND: DIFFUSION MODELS

In this work, the main focus is the image inpainting capability of denoising diffusion models. To outline the background of the employed techniques, a summary of the denoising diffusion generative approach as well as the variants thereof designed for inpainting is provided below.

2.1. Denoising Diffusion Generative Models

Denoising diffusion generative models [17, 8] can overcome some common problems of earlier generative frameworks, such as the mode collapse and unstable trainign in GANs [7, 12] or suboptimal quality of synthesis in VAEs [9].

The generative process of denoising diffusion is based on a gradual transformation of a normal Gaussian sampling distribution $\mathcal{N}(0, I)$ to the distribution of data, by approximating the *reverse process* using a deep neural network. The *reverse process* is modelled as the reverse of the *forward process* that transforms from the data distribution to a Gaussian distribution. The forward chain is obtained by iteratively degrading an image $\mathbf{x}_0$ from the data distribution for T timesteps until reaching $\mathbf{x}_T$. The degradation is performed as additive Gaussian noise using a noise schedule $\beta_0, \beta_1 ... \beta_T$, with β representing the variance of the noise injected into the image:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t, \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

If $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{t=0}^{T} \alpha_t$, the corrupted image $\mathbf{x}_t$ at time step t can be derived as $\mathbf{x}_0$:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{2}$$

The reverse operation cannot be expressed in closed form, and its distribution is instead approximated as $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, t) = \mathcal{N}(\mathbf{x}_{t-1}, \mu_\theta, \tilde{\beta}_t\mathbf{I})$, where:

$$\mu_\theta = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_{t-1} - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\right) \tag{3}$$

The estimation of the noise sample ϵ_θ can be obtained using a deep neural network with parameters θ trained using x_t input and ϵ output derived from Equation 2. As detailed in [8], a reverse step can be computed as:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z} \quad (4)$$

A new sample can be generated by deriving consecutive samples of the diffusion chain, starting from $t = T$ (a sample derived from pure noise), and iterating towards $t = 0$, which arrives at the original data distribution. The exact number of steps T is a hyperparameter, and in the standard definition of the diffusion probabilistic models [8], can be expected to be in the range from several hundred to several thousand. This makes diffusion models generally more computationally expensive than earlier models, like GANs [7] or VAEs [9].

In the domain of Earth Observation, the application of diffusion models has so far been moderate and involved tasks such as cloud removal [19] or controllable image synthesis [18]. The use cases proposed in this work aim to motivate further work on this topic and demonstrate the diversity of potential impact.

2.2. Inpainting with Denoising Diffusion Models

Inpainting can be performed in at least two distinct ways based on diffusion models. One variant, RePaint [10] is based on a masked mixing operation applied to the denoised signal, where an unconditional generative model can be employed for the task. The second variant requires expansion of the denoising network input channels to accommodate the extra condition of the inpainted image and the corresponding mask.

Individual use cases presented in this work demonstrate both variants.

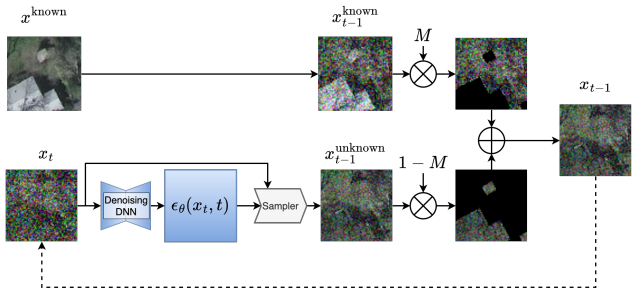

Fig. 1. Inpainting Variant 1: RePaint [10], where a partially known image x^{known} is mixed with the generated sample x_t using a mask M to produce an inpainting at a time step $t-1$.

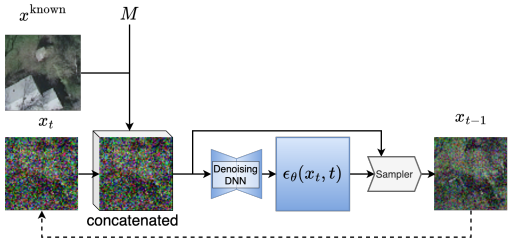

Fig. 2. Diagram of the concatenation inpainting approach for diffusion models.

2.2.1. *Variant 1: RePaint*

The RePaint technique [10] allows for the reuse of an unconditional denoising diffusion model by mixing the denoised generated image with the input masked image x^{known}. As shown in Figure 1, this can be implemented via mixing the generated image x_{t-1} with the noisy input masked image x_{t-1}^{known} (passed through the forward chain until $t-1$) using the mask M at each step of the reverse chain.

In this work, a version of RePaint with modifications is used, where we do not deploy further steps inside every timestep iteration.

2.2.2. *Variant 2: Concatenation*

Another popular approach for diffusion-based inpainting relies on input concatenation, as illustrated in Figure 2, where the denoising network ingests the diffused signal x_t, but also the known image x^{known} and the inpainting mask M. Unlike the RePaint variant described above, this requires a new network architecture to be defined and trained. This increased cost, however, gives an opportunity to create a model optimized directly for the inpainting task. This approach has been commonly used in LatentDiffusion [14] and StableDiffusion [14] models (albeit in latent space, rather than direct image space), and in this work, an inpainting StableDiffusion model is used to generate text-conditioned samples in the urban replanning use case.

3. EXPLORING EO USE-CASES

Three use cases, all taking advantage of the inpainting capability of diffusion models are described below to demonstrate the diversity of potential areas of impact in the field of Earth observation.

3.1. Inpainting for Cloud Removal

The presence of clouds in satellite images poses a challenge both for visual inspection and for downstream tasks such as

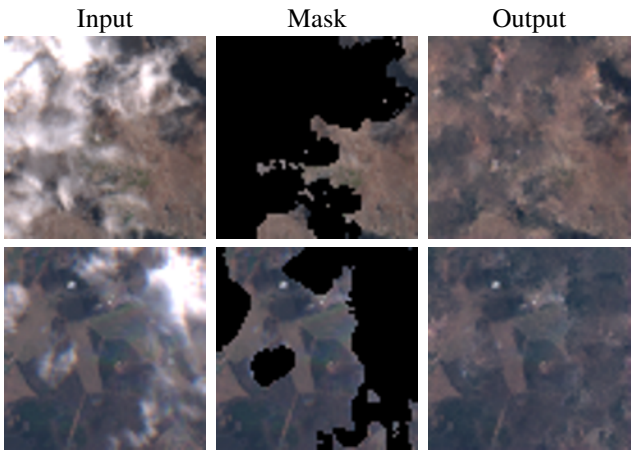

Fig. 3. Selected samples computed on the cloud removal task. Note that these samples of real clouds come from a different set than our reported numerical results

land cover classification. This is particularly problematic for critical real-time applications, where it is not feasible to wait for new acquisitions.

Demonstrated below is how an unconditional diffusion model can be utilised for the task of cloud removal. A custom model has been trained on a set of about 8,000 Sentinel-2 images (RGB bands only in 64x64 resolution) for 250 epochs with 128 of batch size, and then used with the RePaint variant [10] described earlier. This has been tested on the Sentinel-2 Cloud Mask Catalogue [6], a Sentinel-2 dataset containing cloud masks. Applied as an inpainting scheme by reusing random masks applied on the cloud-free images to ensure access to ground truth, the proposed model achieved an SSIM of 0.691 and PSNR of 24.593 dB. Inpainted examples from samples with real clouds are shown in Figure 3 where the dataset cloud masks were used for inpainting. The noise estimator is designed as a 4-layer U-Net with 128 base channels and a channel multiplier factor $f \in \{1, 2, 3, 4\}$.

3.2. Generation of Change Detection Dataset

In EO there is a substantial need to label existing datasets to help current supervised approaches. Moreover, in some applications we have limited datasets in size, which lead to a poor supervised training

Another custom unconditional model has been trained on a remote sensing collection with the aim to generate a new dataset for change detection and consequently, provide more data for training on the downstream task. The model is used to generate a new pair of images, where the actual "change" is generated using our diffusion model applying the inpainting approach we already used for cloud detection. The training is performed for 250 epochs with batch size 128. The neural network is a four-layer U-Net with 128 base channels and a channel multiplier factor $f \in \{1, 2, 3, 4\}$. We train on the whole dataset cropped into 64×64 patches, with 50% overlap between patches.

To ensure the realism of the change regions (they are generally correlated with the overall structure of the image), the original masks and images from OSCD [5] are used and as a result, the original change is replaced with a synthetic one using the trained model. Due to the variational nature of the diffusion process, the resulting inpaintings are likely to introduce some change to the region.

We make this dataset available in open access at the following DOI:10.5281/zenodo.8144237 to motivate its use for supervised change detection pipelines.

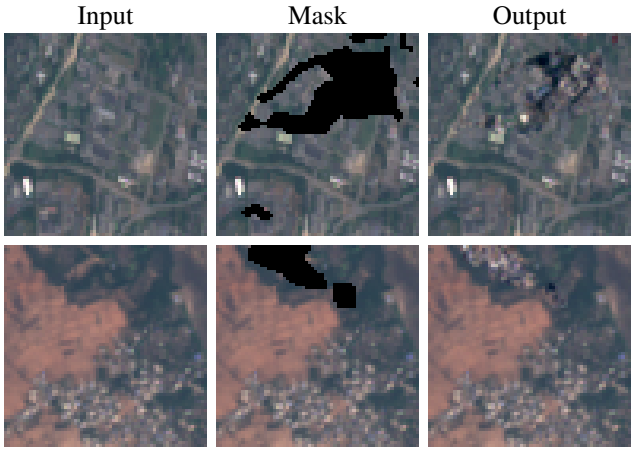

Fig. 4. Example images from the generated change detection dataset.

3.3. Urban Replanning

Another promising use case for the generative models proposed here relates to the urban replanning. With the ability to generate a wide range of realistic images, these models can serve as a visualization tool for communicating ideas about changes to urban environments. The capability to use text as an additional condition provides an interface for the user to control the process and specify the desired content of the generated image.

This use case is demonstrated here by employing a StableDiffusion inpainting variant [14] trained on samples from the LAION-5B dataset (which has been demonstrated to contain a considerable number of Earth observation images [4]) to recreate several scenes from the Inria Aerial Labeling dataset [11], containing high-resolution aerial RGB images (30 cm) of detailed urban scenes. By masking out parts of the existing satellite images, the model can reconstruct alternative versions of the same image based on a provided text-prompt.

Two examples of this use case are shown in Figure 5, where a car park in Vienna (top row) is masked and replaced with a pedestrianised area and a car park in Austin (bottom

row) with a large swimming pool. Transitions between real and generated parts are smooth and realistic thanks to diffusion.

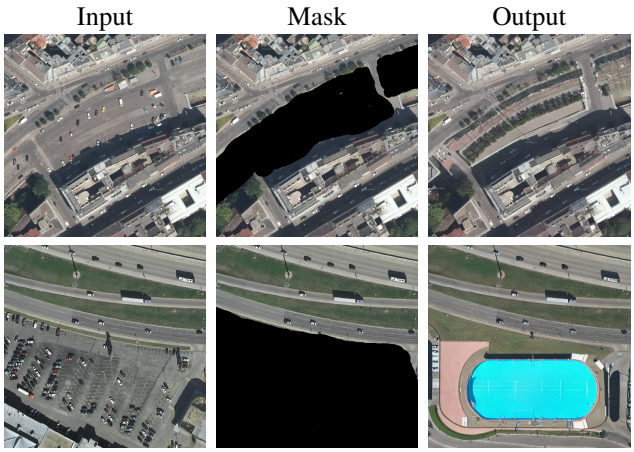

Fig. 5. Example of urban replanning visualization, where a car park is replaced with a pedestrianised area with trees (top row) and another car park is replaced with a large swimming pool (bottom row).

The application of diffusion-based models for urban replanning can lead to reduced costs of visualisation, where the model can support urban designers by generating images representing the potential ideas of rearranging urban spaces. The examples shown here were produced using an off-the-shelf model for image inpainting, which means that these results could potentially be improved further by training specialised models for this task.

4. CONCLUSION

Diffusion models are a promising solution for generative modelling and hence, can support a wide range of use cases in the domain of Earth observation. The three presented examples demonstrate the value of diffusion models for a diverse set of tasks that involve satellite imagery. The inpainting capability of diffusion models has been verified by generating high-quality predictions for the tasks of cloud removal, dataset generation, and urban replanning, with the hope that it inspires the introduction of further remote sensing applications of diffusion models for the benefit of other downstream tasks.

REFERENCES

[1] Nicolas Audebert et al. Deep learning for urban remote sensing. In *JURSE*. IEEE, 2017.

[2] Nicolas Audebert et al. Generative adversarial networks for realistic synthesis of hyperspectral samples. In *IGARSS*. IEEE, 2018.

[3] Javiera Castillo-Navarro et al. Energy-based models in Earth observation: From generation to semisupervised learning. *IEEE Trans. Geosci. Rem. Sens.*, 60, 2021.

[4] Mikolaj Czerkawski and Alistair Francis. From LAION-5B to LAION-EO: Filtering Billions of Images Using Anchor Datasets for Satellite Image Extraction. In *ICCV Workshops*, 2023.

[5] Rodrigo Caye Daudt et al. Urban change detection for multispectral Earth observation using convolutional neural networks. In *IGARSS*. IEEE, 2018.

[6] Alistair Francis et al. Sentinel-2 cloud mask catalogue, 2020. URL https://doi.org/10.5281/zenodo.4172871.

[7] Ian Goodfellow et al. Generative Adversarial Networks. In *NeurIPS*, volume 27, 2014.

[8] Jonathan Ho et al. Denoising diffusion probabilistic models. In *NeurIPS*, volume 33, 2020.

[9] Diederik Kingma et al. Auto-Encoding Variational Bayes. In *ICLR*, 2014.

[10] Andreas Lugmayr et al. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022.

[11] Emmanuel Maggiori et al. Can semantic labeling methods generalize to any city? The INRIA aerial image labeling benchmark. In *IGARSS*. IEEE, 2017.

[12] Gonzalo Mateo-García et al. Generative adversarial networks in the geosciences. In *Deep Learning for the Earth Sciences*, chapter 3. 2021.

[13] Aditya Ramesh et al. Zero-shot text-to-image generation. In *ICML*. PMLR, 2021.

[14] Robin Rombach et al. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[15] Chitwan Saharia et al. Image super-resolution via iterative refinement. *IEEE TPAMI*, 45(4), 2022.

[16] Chitwan Saharia et al. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH Conference*, 2022.

[17] Jascha Sohl-Dickstein et al. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

[18] Zhiqiang Yuan et al. Efficient and controllable remote sensing fake sample generation based on diffusion model. *IEEE Trans. Geosci Rem. Sens.*, 2023.

[19] Xiaohu Zhao et al. Cloud removal in remote sensing using sequential-based diffusion models. *Remote Sensing*, 2023.