Deep Generative AI Based on Denoising Diffusion Probabilistic Models for Applications in Image Processing

Emili Silva Bezerra
emili.bezerra@sou.ufac.br
https://orcid.org/0000-0003-4519-8332
PAVIC Laboratory, University of Acre (UFAC), Brazil

Quefren Oliveira Leher
quefren.leher@sou.ufac.br
https://orcid.org/0009-0005-6678-1131
PAVIC Laboratory, University of Acre (UFAC), Brazil

Uendel Diego da Silva Alves
uendel.alves@sou.ufac.br
https://orcid.org/0009-0009-3357-4979
PAVIC Laboratory, University of Acre (UFAC), Brazil

Thuanne Paixão
thuanne.paixao@sou.ufac.br
https://orcid.org/0000-0002-5563-8971
PAVIC Laboratory, University of Acre (UFAC), Brazil

Ana Beatriz Alvarez
ana.alvarez@ufac.br
https://orcid.org/0000-0003-3403-8261
PAVIC Laboratory, University of Acre (UFAC), Brazil

Received: September 9th, 2024 Accepted: October 8th, 2024
doi: https://doi.org/10.26439/interfases2024.n020.7389

ABSTRACT. Denoising diffusion probabilistic models (DDPMs) have demonstrated significant potential in addressing complex image processing challenges. This paper explores the application of DDPMs in three different areas: reconstruction of remote sensing imagery affected by cloud cover, reconstruction of facial images with occluded areas, and segmentation of bodies of water from remote sensing imagery. Inpainting involves filling in missing regions in images, while DDPMs act as data generators capable of synthesizing information that alings coherently with the context of the original data. Inspired by the inpainting technique, the RePaint approach was adapted and applied to

reconstruction tasks. The WaterSegDiff approach, which uses a diffusion model as a backbone, was employed for the segmentation task. To illustrate the model's behavior and provide examples of the tasks, experiments were carried out with both qualitative and quantitative evaluations. The qualitative results show the model's ability to generate data for reconstruction and segmentation. Quantitatively, metrics such as MSE, PSNR, SSIM, IoU, PA and F1 score highlight the model's proficient performance in image processing tasks. In this scenario, DDPMs have proved to be a promising tool for high-quality data reconstruction, enabling the hallucination of image regions with high visual coherence and facilitating applications in various areas, such as environmental monitoring, facial recognition, water resource mapping, among others.

KEYWORDS: machine learning / reconstruction / segmentation / face and gesture recognition / remote sensing.

IA GENERATIVA PROFUNDA BASADA EN MODELOS DE DIFUSIÓN DE DESENFOQUE PROBABILÍSTICO PARA APLICACIONES EN PROCESAMIENTO DE IMÁGENES

RESUMEN. Los *denoising diffusion probabilistic models* (DDPMs) han mostrado un potencial significativo en la resolución de problemas complejos de procesamiento de imágenes. Este estudio explora el uso de DDPMs en tres aplicaciones diferentes, incluyendo la reconstrucción de imágenes de teledetección en zonas con nubosidad, la reconstrucción de imágenes faciales con regiones ocluidas y la segmentación de masas de agua a partir de imágenes de teledetección. El *inpainting* consiste en rellenar las regiones omitidas en las imágenes, mientras que los DDPM actúan como generadores de datos capaces de sintetizar información coherente con el contexto de los datos originales. En este contexto, tomando la técnica de *inpainting* como inspiración, se adaptó el enfoque *RePaint* y se aplicó a tareas de reconstrucción. Para la tarea de segmentación se utilizó la técnica WaterSegDiff, que también utiliza un modelo de difusión como *backbonner*. Para ilustrar el comportamiento del modelo y ejemplificar las tareas, se realizaron experimentos cuya performance se evaluó cualitativa y cuantitativamente. Los resultados de las evaluaciones cualitativas muestran la capacidad del modelo para generar datos para la reconstrucción y la segmentación. Cuantitativamente, las métricas MSE, PSNR, SSIM, IoU, PA y F1-Score indican un hábil desempeño de los modelos en tareas de procesamiento de imágenes. En este escenario, los DDPMs han demostrado ser una herramienta prometedora para la reconstrucción de datos de alta calidad, permitiendo la alucinación de regiones de imágenes con alta coherencia visual y aplicaciones en diversas áreas, tales como monitoreo ambiental, reconocimiento facial, mapeo de recursos hídricos, entre otros.

PALABRAS CLAVE: aprendizaje automático / reconstrucción / segmentación / reconocimiento facial y gestual / teledetección.

1. INTRODUCTION

In recent years, advancements in deep learning techniques, especially convolutional neural networks (CNNs), have boosted the emergence of artificial intelligence generated content (AIGC). This term refers to data generated by deep learning algorithms, which are capable of creating high-quality content in a variety of formats, such as texts, images and videos. The ability of AIGCs to create content that is nearly indistinguishable from human-made material has revolutionized various sectors, including entertainment, education and scientific research, opening up new possibilities for the content creation and information consumption (Wu et al., 2023; Cao et al., 2023).

According to Zhang et al. (2023), the core of content generation algorithms, also known as generative models, lies in their ability to learn the patterns within a dataset and, based on this knowledge, generate new similar content. Image synthesis is a key area where these models are used to create visually coherent images. Tasks such as super-resolution imaging (enhancing image resolution), image generation from textual descriptions (Text-to-Image) and image reconstruction (Image-to-Image) are examples of diffusion models applications, as illustrated in Figure 1.

Figure 1

Diffusion Model Applications: (a) Image Synthesis for Super-Resolution, (b) Text-to-Image, and (c) Image to Image

Within the field of generative models, diffusion models have proven capable of reversing degradation processes, by learning to recover lost information and generate realistic data (Rombach et al., 2022). These models, also known as denoising diffusion

probabilistic models (DDPMs), often simply referred to as diffusion models for brevity (Ho et al., 2020). DDPMs have demonstrated great potential across various image restoration tasks, including applications in computer vision, robust machine learning, natural language processing, temporal data modeling, multimodal modeling, medical image reconstruction. They have also found interdisciplinary applications in areas such as computational chemistry, image inpainting, image noise removal (denoising), remote sensing, face restoration with occluded areas—particularly in security applications—and image segmentation (Yang et al., 2023).

In remote sensing applications, Singh and Vyas (2022) emphasize the high homogeneity and geospatial accuracy of the data obtained by remote sensing, while underscoring the potential for occlusions in adverse conditions. These occlusions, often associated with cloud cover, can impair the quality of vegetation indices, which are mathematical models used to quantify characteristics of the Earth's surface. One notable example is land surface temperature (LST), an indicator sensitive to changes in resource and environmental conditions, especially in areas with high spatio-temporal variability. According to Awais et al. (2022), LST is influenced by multiple factors, such as human activities as well as vegetation and soil water conditions. García and Díaz (2021) further corroborate the importance of LST across various areas of knowledge, including hydrology, meteorology, surface energy balance and climate studies. Growing concern about the effects of climate change has led to the identification of certain ecosystems as key indicators of environmental impact. Lakes, for instance, are often sensitive and rapid sentinels of climate and hydrological changes in river basins, providing valuable tools for understanding environmental dynamics (Adrian et al. 2009). For example, Perez-Torres et al. (2024) developed automated and efficient methods to accurately capture lakes in high mountain environments, aiming to climate change challenges in these ecosystems, prevent and mitigate disasters and properly manage and protect water resources.

Simultaneously, the prevalence of images in modern society makes image restoration a critical research area in computer vision. The presence of artifacts, noise, or missing regions in images can compromise the quality of visual information and human interpretation. Inpainting, an image processing technique, aims to fill in these gaps coherently with the visual context, significantly improving the perceptual quality of images (Elharrouss et al., 2020; Li et al., 2023). Facial recognition, an intuitive task for humans, poses complex challenges for computational systems. Changes in capture conditions, such as lighting and angle, as well as individual variations, significantly impact the accuracy of algorithms (Kortli et al., 2020). The primary objective of facial recognition systems is to identify individuals from static images or video sequences (Ali et al. 2021; Taskiran et al., 2020). To deal with degraded images or occluded images, inpainting—especially when based on diffusion models—is emerging as a promising technique. This approach shows potential for improving the robustness of facial recognition systems, especially in challenging conditions.

Based on the works by Leher (2024), Alves (2024) and Perez-Torres et al. (2024), this paper presents the versatility of DDPMs. Expanding on the research by Lugmayr et al. (2022), it refines using DDPMs to recover lost information more precisely across various image modalities. Furthermore, integrating a segmentation model with DDPMs facilitates the extraction of bodies of water, this approach in multiple applications, opening up new possibilities for visual data analysis and interpretation.

2. RELATED STUDIES

Sohl-Dickstein et al. (2015) pioneered the application of diffusion models, introducing a new method for modelling complex data. They proposed a technique involving the gradual destruction of the data structure, followed by learning an inverse process for its reconstruction. This method resulted in a deep and versatile generative model that enabled rapid learning, efficient sampling and precise calculation of probabilities. Inspired by statistical physics, the approach offers a solution to the challenge of balancing flexibility and tractability in data modelling.

Various studies have since focused on diffusion model-based for image reconstruction. Avrahami and Fried (2022), in the work *Blended Diffusion for Text-driven Editing of Natural Images*, presented an inpainting solution to perform local (region-based) editing on generic natural images based on a natural language description together with a region of interest (ROI) mask. They combined a contrastive language-image pretrained (CLIP) model to direct editing to a user-supplied text prompt with a DDPM to generate natural-looking results.

Similarly, Lugmayr et al. (2022), in *RePaint: Inpainting using Denoising Diffusion Probabilistic Models*, proposed an inpainting approach based on a DDPM-type diffusion model, achieving high-quality results even with atypical masks. They used an unconditional DDPM pre-trained with generative priors, altering the reverse diffusion iterations by sampling the unmasked regions using the provided image information. As this technique does not modify or condition the original DDPM network itself, the model produces diverse, high-quality output images for any form of inpainting/filling.

Kawar et al. (2022) introduced denoising diffusion restoration models (DDRMs), a diffusion model based on an unsupervised posterior sampling method, achieving efficient results in the fields of image restoration, super-resolution, deblurring, colorization and inpainting. DDRM proved to be an excellent solver of linear inverse problems through general sampling with an unconditional diffusion model.

Approaches based on diffusion models have proven effective in generating new data, which has motivated recent research into their application in remote sensing imagery. Liu et al. (2022) and Bandara et al. (2022) explored these models in different contexts. Liu et al. (2022) proposed the diffusion model with detail complement (DMDC),

a generative model specifically for super-resolution images. Rather than just optimizing existing images, DMDC generates high-resolution images from low-resolution inputs, allowing for a deeper understanding of the image and the recovery of fine and complex details that could be lost at lower resolutions. On the other hand, Bandara et al. (2022) applied DDPMs for feature extraction, to improve the accuracy of change detection in remote sensing images. The model, trained on a large set of unlabeled images, learned to generate images from noise. After training, DDPMs are able to extract relevant features such as texture, shape, and patterns, which were used to train a simple classifier that identifies changes in specific image areas.

In their study on cloud removal from satellite images, Jing et al. (2023) introduced the DDPM-CR model, which is notable for its ability to remove both thin and thick clouds from radar data. This model, based on the DDPM architecture, uses the SEN12MS-CR database to improve its results. DDPM-CR integrates cloud-contaminated optical images with synthetic aperture radar (SAR) images, where the information provided by the SAR images helps to accurately reconstruct the areas obscured by the clouds. The model incorporates a multi-scale attention mechanism for effective cloud identification and removal. In addition, the loss function developed for training the model is specifically designed for cloud removal, considering both high- and low-frequency information.

In parallel, Zhao and Ji (2023) proposed the sequential-based diffusion models (SeqDMs), which combine data from different sources, such as radar images (unaffected by clouds) and optical images (interfered by clouds). SeqDMs analyze temporal sequences of images to generate more precise information about cloud-covered areas. The model is adaptable to sequences of different lengths, which enhances its applicability across different situations.

Inspired by recent advancements in natural language processing, computer vision, and image synthesis from Gaussian noise, diffusion probabilistic models and transformer models have demonstrated a remarkable ability to capture complex spatial and contextual relationships, generating high-quality images. This makes them particularly suitable for image segmentation tasks. While traditionally used for image generation and inpainting, diffusion models have recently been applied to semantic segmentation. Amit et al. (2021) introduced an innovative technique that integrates the power of diffusion models, known for their ability to generate high-quality images, with image segmentation task. Unlike approaches that rely on pre-trained models in other tasks (backbones), their model is trained in an integrated way, combining the information from the original image with the current segmentation estimate by means of two encoders. Through additional coding layers and a decoder, the model iteratively refines the segmentation, using the probabilistic mechanism characteristic of diffusion models.

Semantic segmentation models, which divide images into different semantic regions, have difficulty identifying the exact boundaries between these regions. This

challenge arises because convolutional operators, common tools in these models, tend to smooth out fine details, making it difficult to clearly distinguish the boundaries. Tan et al. (2022) proposed a new technique called semantic diffusion network (SDN) to improve the ability of segmentation models to detect boundaries. SDN works as an anisotropic diffusion process that emphasizes edge and texture information relevant to semantic boundaries. SDN creates a mathematical mapping that transforms the original features into boundary-sensitive features.

Ayala et al. (2023) proposed a solution for semantic segmentation in remote sensing imagery conditioning the diffusion process to reduce noise in input images. This method aims to guide the generation of the segmentation mask, ensuring consistency with the elements present in the original image. By conditioning the diffusion process, the authors significantly improved the mask's accuracy, so that it adequately reflects the elements of the captured scene. The model was evaluated on a specific aerial images dataset and compared to state-of-the-art techniques. The results demonstrate the promising potential of this approach as a valuable tool in the field of remote sensing.

3. METHODOLOGY

3.1 DDPM

As outlined by Dhariwal and Nichol (2021), the diffusion process consists of two primary stages: forward diffusion and reverse diffusion.

3.2.1 Forward diffusion

In this stage, a Markov chain is used to progressively add small amounts of noise ε to a data sample x_0, producing a sequence of increasingly noisy samples $x_0, x_1, \ldots, x_T$. The amount of noise at each step is controlled by the variance $\{\beta_T \in (0,1)\}_{t=1}^{T}\}$, as seen in Equation 1.

$$q(x_t|x_{t-1}) = \aleph(x_t \sqrt{1-\beta_t}x_{t-1}, \beta_t I \qquad (1)$$

As increases, the original data sample x_0 loses its distinctive characteristics and eventually becomes an isotropic Gaussian distribution x_T.

3.2.1 Reverse diffusion

In the reverse diffusion stage, the process uses a neural network with parameters denoted as θ to iteratively remove noise from the noisy sample x_T at each time step until a high-quality sample x_0 is obtained. A neural network predicts the mean $\mu_\theta(x_t, t)$ and variance $\Sigma_\theta(x_t, t)$ of the Gaussian distribution, calculated with Equation 2:

$$p_\theta(x_{t-1}|x_t) = \aleph(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (2)$$

The learning objective of the model is derived by considering the variational lower bound (VLB) between the prior and posterior distributions, as seen in Equation 3:

$$L_{VLB} = E_q\Big[D_{KL}\big(q(x_t|x_0)||p_\theta(x_T)\big) + \sum_{t>1} \Big(D_{KL}\big(q(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)\big)\Big) - logp_\theta(x_0|x_t)\Big]$$

$$L_{VLB} = E_q[L_T + \sum_{t>1}(L_{t-1}) - L_0] \quad (3)$$

After a series of derivations of the terms L_{t-1},, one simplified training objective L_{simple} is obtained, according to Equation 4:

$$L_{simple} = E_{t\sim[1,T],x_0,\varepsilon}[\,||\varepsilon_t - \varepsilon_\theta(x_t, t)||^2\,] \quad (4)$$

Since does not depend on the variance $\Sigma_\theta(x_t, t)$, a new hybrid object is defined according to Equation 5:

$$L_{hybrid} = L_{simple} + \lambda L_{VLB} \quad (5)$$

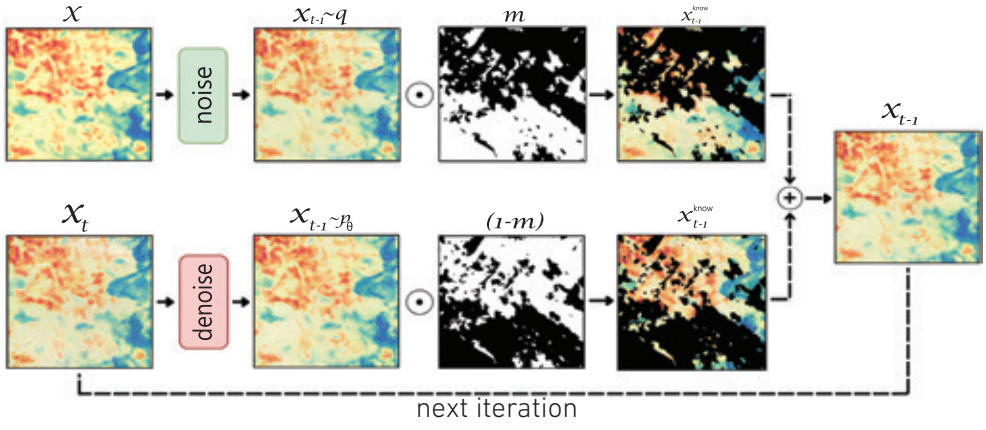
3.2 Image Reconstruction

After the consolidation of diffusion models (DMs), the inpainting task has been significantly optimized. Studies by Sohl-Dickstein et al. (2015) and Ho et al. (2020) demonstrated the potential of diffusion models. Since then, several authors, including Lugmayr et al. (2022), have introduced new frontiers in computational inpainting using diffusion models, achieving outstanding results.

Lugmayr et al. (2022) propose a robust free-form inpainting method called RePaint, which fills arbitrary regions of an image defined by a mask. The method uses a pre-trained DDPM-type (Ho et al., 2020), conditioning the generation process only in the reverse diffusion iterations by sampling the unmasked regions. Figure 2 illustrates RePaint's iterative process. The methodology proposes a reformulation of the traditional denoising process, aimed at conditioning the content of the input image. In each iteration, samples from the known region of the original image (upper sequence in Figure 2) and the already filled portion of the output generated by the DDPM (lower sequence in Figure 2) are used as input for the model. Preserving the DDPM's original architecture guarantees the diversity and high quality of the generated images, regardless of the shape of the inpainting mask. This feature gives the technique greater flexibility, allowing arbitrary masks to be applied and inpainting to be carried out more freely.

Figure 2

RePaint Process that Exemplifying the Reconstruction of an LST Image



Note. From "RePaint: Inpainting using denoising diffusion probabilistics models" by A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, & L. Van Gool, 2022, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11461-11471 (https://openaccess.thecvf.com/content/CVPR2022/papers/Lugmayr_RePaint_Inpainting_Using_Denoising_Diffusion_Probabilistic_Models_CVPR_2022_paper.pdf).

Considering Figure 2, the original image (ground truth) is designated by x, the unknown pixels by $m \odot x$ and the known pixels by $(1 - m) \odot x$, where m represents the original mask region and $(1 - m)$ represents the inverse mask region. Therefore, the known regions x_{t-1}^{unknow} (unmasked pixels) are represented by the values 1 in the binary mask matrix , indicating an unmasked pixel, while the value 0 indicates a pixel covered by the mask. Thus, x_{t-1}^{unknow} is sampled from the input data by Equation 6:

$$x_{t-1}^{know} \sim \mathcal{N}(\sqrt{\underline{\alpha_t}} x(1 - \underline{\alpha_t})I) \qquad (6)$$

In this way, the unknown regions x_{t-1}^{unknow} (covered by the mask) by sampling from the neural model, shown in Equation 7:

$$x_{t-1}^{unknow} \sim \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \qquad (7)$$

Finally, these samples are combined to obtain the next reverse intermediate step x_{t-1} , as shown in Equation 8:

$$m \odot x_{t-1}^{know} + (1 - m) \odot x_{t-1}^{unknow} \qquad (8)$$

In this way, reconstruction of the occluded areas after applying the mask is carried out using the inpainting technique, generating a reconstructed image in the output. By

definition, the area to be reconstructed is delimited by a binary mask m, which is filled in based on information from the surrounding pixels according to the pre-trained DDPM.

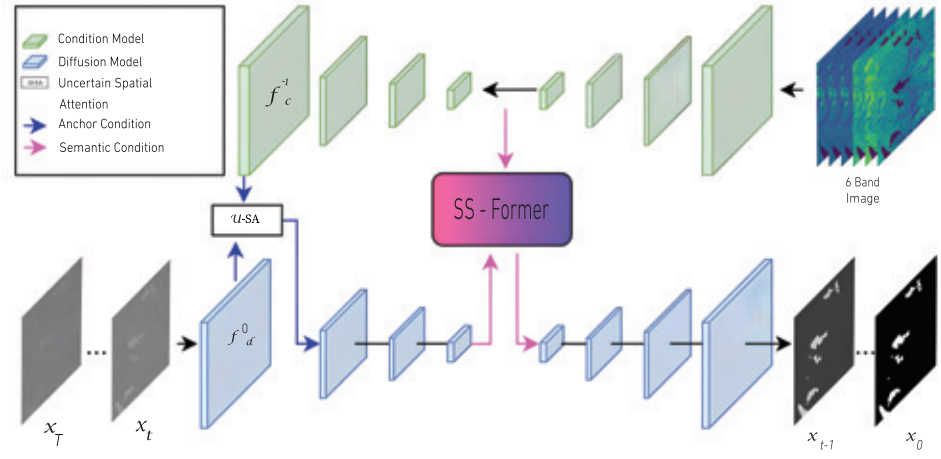
3.3 Image Segmentation

Diffusion probabilistic models with noise reduction have shown remarkable results in generative image modeling. Studies by Amit et al. (2021), Tan et al. (2022), and Ayala et al. (2023) highlight the potential of these models in semantic segmentation tasks, particularly in remote sensing. However, the main challenge in applying them arises from their generative nature, which creates a segmentation mask from random noise. To ensure that this mask corresponds to the target image, a restricted diffusion process is needed to guide the generation of the mask in a more accurate and coherent way.

MedSegDiff-V2 (Wu et al., 2023) is a transformer-based diffusion framework that uses two different conditioning techniques—anchor condition and semantic—condition which effectively integrate the conditioning resources into the diffusion model. This framework employs two U-Net architectures: one for the diffusion block and one another for the conditioning block. The condition U-Net block acts as a segmentation feature extractor from the original raw image, learning the most relevant features. These segmentation features are integrated with the noise mask information using the anchor condition technique, which implements the uncertain spatial attention (U-SA) mechanism. The integrated data is fed into the U-Net diffusion model's encoder. On the other hand, the semantic condition integrates the high-level features obtained by the diffusion and conditioning models through a transformation mechanism called spectrum-space transformer (SS-Former). This cross-attention mechanism operates in the frequency domain, aligning the noisy image data with the segmentation features of the raw image. Both conditioning mechanisms address the incompatibility issue of combining a U-Net model with diffusion probabilistic models, by implementing an interface between the two models, which helps to reduce the large variations in the transformer configuration. Based on MedSegDiff, a new architecture called WaterSegDiff has been developed specifically for the task of segmenting bodies of water from remote sensing imagery. Figure 3 illustrates the WaterSegDiff architecture.

Figure 3

WaterSegDiff Architecture Showing the Two Conditioning Mechanisms: U-SA and SS-Former



Note. From "Exploratory analysis using deep learning for water-body segmentation of Peru's high-mountain remote sensing images," W. I. Perez-Torres, D. A. Uman-Flores, A. B. Quispe-Quispe, F. Palomino-Quispe, E. Bezerra, Q. Leher, T. Paixão, and A. B. Alvarez, 2024, *Sensors*, *24*(16), Article 5177 (https://doi.org/10.3390/s24165177).

At each stage t of the diffusion process, a noisy mask x_t is introduced into the diffusion model. This model is conditioned by segmentation features extracted from the raw image using the conditioning model. The diffusion process is instructed using the Anchor Condition and Semantic Condition techniques, where the former allows the diffusion model to be initialised with an approximate but static reference, which helps to reduce variations in diffusion. While the second technique, using the SS-Former, connects the noise and high-level segmentation information to be fed into the diffusion model decoder, generating a more robust representation by taking advantage of the global and dynamic nature of the Transformer proposed by Muzammal et al. (2021). This conditioning introduced into the diffusion model decoder can be expressed as in Equation 9.

$$\epsilon_\theta(x_t, I, t) = D(TransF(E_t^I, E_t^x), t) \qquad (9)$$

Where E_t^I represents the high-level features of the raw image and E_t^x represents the high-level features of the image with current noise. Using a transformer, both features are incorporated and passed through the D decoder of the diffusion model.

4. EXPERIMENTS

The experiments were carried out at the Federal University of Acre (UFAC) Pesquisa Aplicada em Visão e Inteligência Computacional (PAVIC) laboratory in Brazil, encompassing three independent experiments. All experiments were conducted using PyTorch 2.0.1 on Ubuntu 22.04.3 LTS. Section 4.1 presents a methodology for reconstructing remote sensing imagery in order to estimate Earth's surface temperature. Section 4.2 describes image reconstruction techniques for facial recognition applications, security systems and facial biometrics. Finally, Section 4.3 includes a study on the segmentation of bodies of water in satellite images for remote sensing.

4.1 First Experiment: LST Image Reconstruction

The reconstruction of satellite images is a complex problem that requires high computational power and faces challenges such as noise and cloud cover. To mitigate these effects, image preprocessing techniques inspired by Bezerra et al. (2023) were used as a first step in applying the RePaint model.

The experiment conducted by Leher (2024) explores the applicability of an inpainting approach based on DDPMs (Ho et al., 2020) to reconstruct satellite information (Landsat-8) for the calculation of LST in areas affected by cloud cover. This experiment follows the methodology of the surface energy balance algorithm for land (SEBAL) model (Bastiaanssen et al., 1998) to estimate LST. In this process, emissivity is derived from vegetation indices calculated using the reflectance values extracted from the bands B4 and B5. At the same time, spectral radiance is obtained from the band B10. Finally, the information gathered from the emissivity and spectral radiance is used to reconstruct the LST in each image fragment. Thus, the LST is estimated using the following procedures:

- Spectral radiance: Converts digital numbers (DN) from the thermal band (B10) into spectral radiance using band-specific calibration factors.
- Top-of-atmosphere reflectance: Calculates reflectance for bands B4 and B5 using solar elevation, sun-earth distance and band-specific calibration factors.
- Vegetation indices: Derive the normalized difference vegetation index (NDVI) and the soil adjusted vegetation index (SAVI) from reflectance values. The leaf area index (LAI) is calculated from the SAVI and represents the ratio of the leaf area of a vegetation and the area of the unit covered by that vegetation. These indices represent the health and biomass of the vegetation.
- Emissivity: Estimates pixel emissivity using a linear relationship with the SAVI, characterizing the heat radiation properties of a surface.

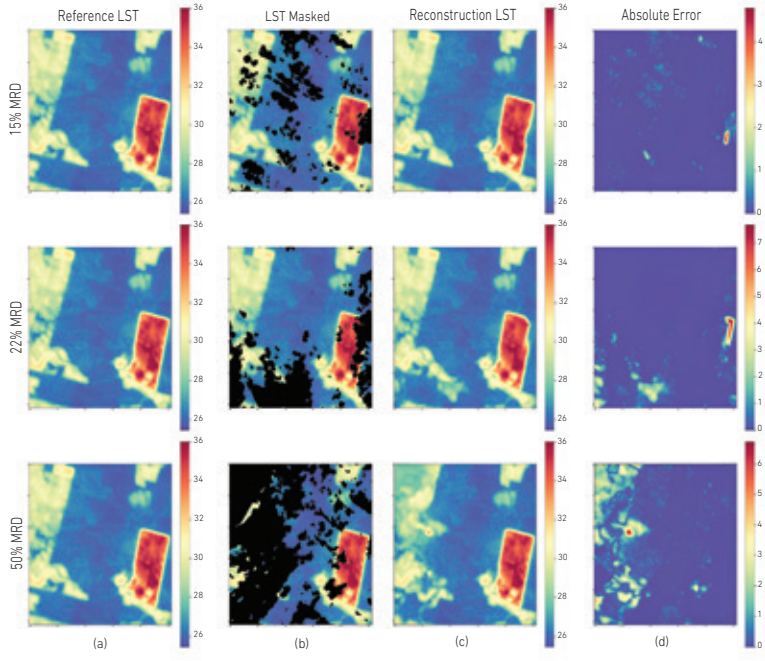
The study region selected for this research is the western Brazilian Amazon rainforest, known for its high annual cloud cover. Figure 4 shows the visual results of the

model reconstructing at three levels of cloud cover: 15%, 22% and 50% missing rate data (MRD).

The complexity of generative models such as RePaint results in significant computational cost. In this study, training on a set of satellite images required 48 hours of processing on dedicated hardware. Training and testing were performed using a 3.0 GHz Intel Xeon Gold 6342 CPU and an NVIDIA HGX A100 GPU. The inference phase - the process of generating new images from the trained model - took an average of 15 minutes per image.

Figure 4

Visual Results of the LST Retrieval: (a) Ground Truth LST, (b) Masked LST, (c) Reconstructed LST, (d) Absolute Error



As shown in Figure 4, the model exhibited strong ability in reconstructing LST images with high fidelity, even under significant cloud cover. The reconstruction with 15% cloud cover achieved the highest quality, with minimal absolute errors compared to the reference LST image. This high quality is evident by both the visual analysis and error quantification, which did not exceed 4°C. Reconstructions for images with 22% and 50% cloud cover also yielded satisfactory results, with maximum errors of 7°C and 6°C respectively. The visual comparison of the reconstructed images against the reference images, as well as absolute error quantification, enabled a robust assessment of the model's performance across cloud cover scenarios.

In addition to the visual analysis of the reconstruction, a quantitative evaluation was carried out using metrics such as mean square error (MSE), for fidelity to the original image, peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM). The numerical results of these metrics are shown in Table 1.

Table 1

MSE, PSNR, and SSIM Quantitative Statistics Values from LST Image Reconstruction

Scenarios	MSE ↓	PSNR ↑	SSIM ↑
15 %	0,0019	47,1894	0,9874
22 %	0,0001	39,9359	0,9763
50 %	0,0002	35,9113	0,9478

Note. ↓ indicates that a lower value is better, while ↑ indicates that a higher value is better.

As expected, the results in Table 1 reveal that cloud cover directly affects the quality of the model's reconstruction. The scenario with 15% cloud cover showed the best results, with the highest PSNR (47,1894) and SSIM (0,9874) metrics, indicating a reconstruction that was more faithful to the original image. The 22% cloud cover scenario, despite achieving the lowest MSE (0,0001), obtained slightly lower results in the other metrics. The 50% coverage scenario yielded the lowest PSNR and SSIM values among the three scenarios. However, even at this level of cloud cover, the metric values indicate good reconstruction quality, suggesting the robustness of the model across different cloud cover conditions.

4.2 Second Experiment: Reconstruction of Facial Images

The RePaint model was applied to a set of facial images from the CelebA-HQ dataset, following the pre-processing technique proposed by Alves (2024). This stage, which involves extracting the facial region of interest (ROI), aims at guiding the model's learning toward facial patterns, thereby improving its reconstruction ability. This research uses RePaint with transfer learning, and this pre-processing stage has proven crucial for increasing the robustness of facial recognition systems in scenarios with partial occlusion.

This experiment introduces a dynamic approach using the inpainting technique based on diffusion models to reconstruct occluded areas of facial images. From a single sample, the model realistically synthesizes several missing parts, enabling the extraction of facial features critical for identification. Figure 5 displays the reconstruction results for two randomly chosen samples.

The computational cost of generative models such as RePaint pose a significant challenge, particularly in large-scale projects. In this experiment, training on facial images required 120 hours of processing time (Lugmayr et al., 2022), while generating each reconstructed image took approximately 14 minutes and 30 seconds. These findings underscore the necessity for optimizations and the development of more efficient computational infrastructures to facilitate the large-scale implementation of these models.

The qualitative evaluation of facial reconstruction results, shown in Figure 5, indicates that the model effectively recovers lost facial information in low and moderate occlusion scenarios (15% and 20%). Sample 1 (male face) achieved the best reconstruction for 15% occlusion, while sample 2 (female face) visually produced the best results for 20% occlusion. However, the model's performance degrades notably under conditions of high occlusion (50%), highlighting the need for further research to improve the model's robustness in scenarios with more extensive information loss.

Table 2 presents the quantitative statistics for MSE, PSNR and SSIM metrics. The analysis of sample 1 reveals an inverse correlation between the level of occlusion and the quality of the reconstruction. The 15 % occlusion scenario showed the highest MSE, indicating greater fidelity to the original image. This observation is corroborated by the PSNR and SSIM values, which also achieved the best results in the same scenario. The expected gradual decrease in these metrics with increasing reflects the added challenge larger occluded areas pose to the reconstruction process. However, the results demonstrate the model's robustness to generate high-quality reconstructions, even under challenging conditions.

Figure 5

Reconstruction of Two Samples Under 15%, 20% and 50% MRD Occlusion Scenarios. (a) Sample 1, (b) Sample 2

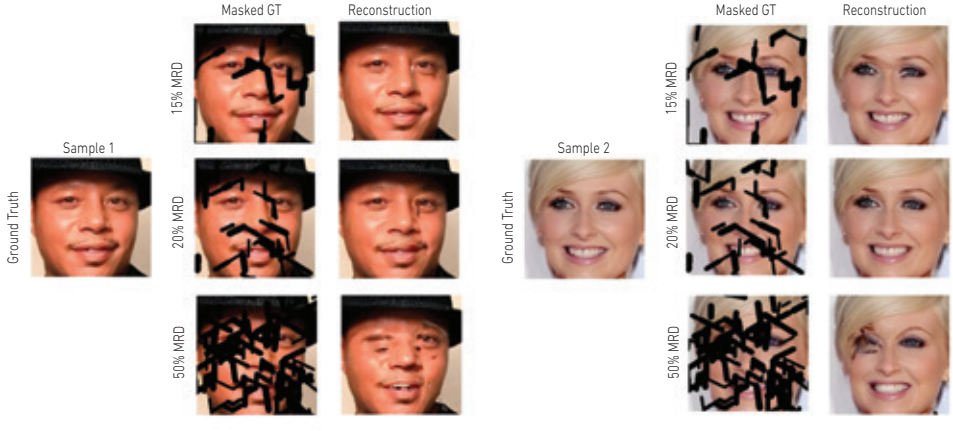


Table 2

MSE, PSNR, and SSIM Quantitative Statistics from Sample 1

Scenarios	MSE ↓	PSNR ↑	SSIM ↑
15 %	0,0044	29,5741	0,9295
20 %	0,0412	29,8718	0,9126
50 %	0,0214	22,7156	0,7773

Note. ↓ indicates that a lower value is better, while ↑ indicates that a higher value is better.

The quantitative statistics of sample 2, presented in Table 3 and using the MSE, PSNR, and SSIM metrics, reveals a different behavior compared to sample 1. The 20% occlusion scenarios showed the lowest MSE, indicating greater fidelity to the original image. This observation is corroborated by the PSNR values which also achieved the best results in the same scenario. However, SSIM reached its peak at 15% occlusion. Similarly to sample 1, the results for sample 2 highlights the model's robustness to generate high-quality reconstructions, even under challenging occlusion conditions.

Table 3

MSE, PSNR, and SSIM Quantitative Statistics from Sample 2

Scenarios	MSE ↓	PSNR ↑	SSIM ↑
15 %	0,0017	33,6736	0,9379
20 %	0,0014	34,6582	0,9291
50 %	0,0123	25,1303	0,8106

Note. ↓ indicates that a lower value is better, while ↑ indicates that a higher value is better.

4.3 Third Experiment: Segmentation of Bodies Water

The accurate segmentation of lakes in high-resolution images, especially in complex mountainous regions such as the Peruvian Andes, presents significant challenges for environmental monitoring and water resource management. The experiment conducted by Perez-Torres et al. (2024) proposes an innovative approach called WaterSegDiff, which is based on diffusion probabilistic models and transformers. WaterSegDiff incorporates semantic anchoring and conditioning mechanisms to capture the distinctive characteristics of lakes across different environmental contexts. Applied to high-resolution images of the Peruvian Andes, WaterSegDiff generates accurate and up-to-date lake maps, essential for monitoring the dynamics of these ecosystems that are sensitive to climate change and land use. The high temporal and spatial frequency of remote sensing data makes it possible to detect changes in the lake area and morphology, providing crucial information for integrated water resources management and Andean biodiversity conservation.

Figure 6 shows three examples of lake segmentation using WaterSegDiff, illustrating different areas covered by bodies of water. The qualitative results demonstrate the model's ability to approximate ground truth, indicating high performance in segmenting bodies of water, even in complex scenarios. Although small discrepancies were observed between the ground truth and predict masks, the model exhibits accurate and robust segmentation, underscoring its potential for applications in hydrological studies and water resource management.

The computational cost of training and running inference on generative models, such as Repaint, while generating a single image after training required around 15 minutes for segmentation processing. Training and testing were performed using a 3.0 GHz Intel Xeon Gold 6342 CPU and an NVIDIA HGX A100 GPU.

To quantify the segmentation ability, the model was subjected to a quantitative evaluation using the intersection over union (IoU), pixel accuracy (PA) and F1 score metrics, as shown in Table 4. Analysis of Table 4 reveals that scenario 2 has the highest IoU value, indicating a better overlap between the segmented regions and reference regions. On the other hand, scenario 1 obtained the best PA and F1 score values, demonstrating high pixel classification accuracy and a good balance between accuracy and suppression. With all metrics exceeding 0,95, these results suggest that the model exhibits strong generalization power for segmentation tasks, even under challenging scenarios.

Figure 6

Segmentation Using WaterSegDiff: (a) showing the RGB Image, (b) Ground Truth Mask, and (c) Predicted Mask

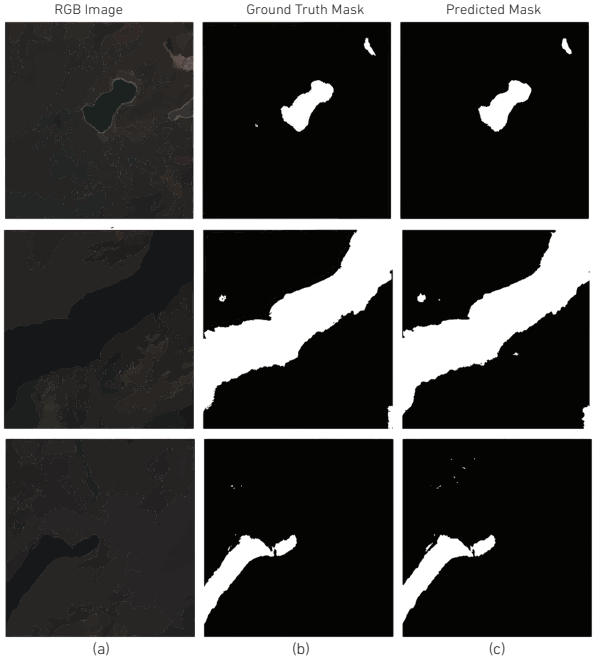


Table 4

Metrics and Standard Parameters for Quantitative Statistics from the WaterSegDiff Model

Scenarios	IoU ↑	PA ↑	F1 Score ↑
1	0,9561	0,9984	0,9776
2	0,9677	0,9900	0,9836
3	0,9502	0,9971	0,9745

Note. ↓ indicates that a lower value is better, while ↑ indicates that a higher value is better.

5. DISCUSSION

The analysis of Figure 4 and Table 1 - i. e., the results of the first experiment - reveals that cloud cover has a significant impact on the quality of LST image reconstruction in Western Amazonia. As cloud cover increases, the MSE, PSNR and SSIM quantitative metrics indicate a progressive deterioration in reconstruction quality, corroborating the findings from the visual analysis. However, the model shows robustness across different cloud cover conditions, yielding promising results even under high cloudiness. The 15% cloud cover scenario achieved excellent results, with the highest PSNR and SSIM values indicating a reconstruction that was both accurate and closely aligned to the original image. As cloud cover increased to 22% and 50%, a gradual degradation in reconstruction quality was observed, although metric values remained satisfactory, especially considering the complexity of the task. These results suggest that cloud cover is the main factor affecting LST image reconstruction quality in the study region. While other factors - such as cloud type, spatial and temporal resolution, and atmospheric conditions - may contribute to variability in results, cloud cover's influence stands out.

Regarding the second experiment, the qualitative results shown in Figure 5 indicate that robust model performs well in reconstructing facial images with low and moderate occlusion levels, though performance degrades significantly under conditions of high occlusion. Quantitative metrics from Tables 2 and 3, corroborate these observations, with MSE, PSNR, and SSIM metrics showing an inverse correlation between the occlusion level and the reconstruction quality. This suggests that the model performs better in scenarios under a lower level of occlusion. The experiment demonstrates the effectiveness of the inpainting technique based on diffusion models for the reconstruction of partially occluded facial images, highlighting its potential for applications in facial recognition systems.

Finally, in the third experiment, the results obtained with the WaterSegDiff model demonstrate high accuracy in segmenting lakes, even in complex scenarios. The comparison between model-generated masks with reference masks ground truth indicates

excellent agreement, with only minor discrepancies in some cases, as shown in Figure 6. The IoU, PA, and F1 score metrics in Table 4 further quantify the model's performance, with all exceeding 0,95 across all scenarios, which suggests a strong generalization power for segmentation tasks. The application of the WaterSegDiff to high-resolution images of the Peruvian Andes enables the generation of accurate and up-to-date lake maps, essential for monitoring the dynamics of these ecosystems that are sensitive to climate change and land use.

While the results demonstrate considerable potential, the computational cost is significantly higher than that of state-of-the-art models in the literature. Notably, the facial reconstruction model required the most extended training time compared to the other two experiments, highlighting the need for optimization processes to make these models more practical for real-time scenarios.

6. CONCLUSIONS

Diffusion probabilistic models have demonstrated significant potential across various image processing tasks, especially for reconstructing occluded or damaged areas and sementing image. This study explored the application of DDPMs in three specific areas: reconstruction of remote sensing imagery for estimating LST in areas with cloud cover, segmentation of bodies of water in satellite images for delimiting areas covered by lakes and reconstruction of facial images with occluded areas, to improving facial recognition systems. Reconstruction was achieved through inpainting where DDPMs enabled the recovery of lost information for more precise and complete analyses. The study also presents two applications of RePaint for reconstruction tasks and WaterSegdiff, which uses a DDPM backbone, transformers, semantic anchoring and conditioning mechanisms, for segmentation.

A notable advantage of diffusion-based approaches, as observed in the experiments, is that these methods require only a single sample to perform reconstruction and/or segmentation, unlike other approaches that need multiple images to complement the task.

In the first experiment, cloud cover emerged as the main factor affecting LST image reconstruction quality in Western Amazonia. Although the model exhibited spatial discontinuities relative visual structures in some areas high cloudiness did not compromise reconstruction accuracy, indicating optimal performance. In the second experiment, the model proved robustness in reconstructing facial images with low and moderate occlusion, though performance degraded under high occlusion. Finally, in the third experiment, WaterSegDiff showed high performance in segmenting lakes, even in complex scenarios, achieving quantitative statistics exceeding 0,95 across all scenarios. Thus, despite challenges from atmospheric conditions, occlusion level, and scene complexity, these

findings demonstrate the strong performance of diffusion models in handling complex tasks such as image reconstruction and segmentation.

Generative diffusion models thus offer an innovative approach with significant potential for intelligent digital image processing applications, including environmental monitoring, facial recognition, and water resource mapping, emphasizing the versatility of diffusion model's.

Although the outcomes are promising, the models utilized in this research present a key limitation: high computational cost. Pursuing optimization strategies is essential to reduce computational time and enhance the efficiency of these models for practical applications. Future research should focus on strengthening the model's resilience in challenging scenarios, particularly by improving their ability to cope with elevated occlusion rates, which are typical of dynamic and complex environments. Furthermore, accurately segmenting more complex environmental elements, such as transparent objects or those with a similar texture to the background, remains a significant challenge to be addressed. These considerations reveal a significant scope for further research and development, as the field faces numerous challenging and complex issues that require in-depth research.

The exploration of new applications for this promising technology offers a rich field for future research. The authors are particularly interested in utilizing generative diffusion models for tasks such as high dynamic range (HDR) image enhancement and generative data augmentation, aimed at improving computer vision capabilities across various domains. Additionally, advancements of quantization, pruning and knowledge distillation techniques hold great promise for reducing model size and accelerating inference, thereby enhancing their suitability for resource-limited devices.

REFERENCES

Adrian, R., O'Reilly, C. M., Zagarese, H., Baines, S. B., Hessen, D. O., Keller, W., David M. Livingstone, D. M., Sommaruga, R., Straile, D., Van Donk, E., Weyhenmeyer, G. A., & Winder, M. (2009). Lakes as sentinels of climate change. *Limnology and oceanography*, *54*(6 part 2), 2283-2297. https://doi.org/10.4319/lo.2009.54.6_part_2.2283

Ali, W., Tian, W., Din, S. U., Iradukunda, D., & Khan, A. A. (2021). Classical and modern face recognition approaches: a complete review. *Multimedia tools and applications*, *80*, 4825-4880. https://doi.org/10.1007/s11042-020-09850-1

Alves, U. D. S. (2024). *Reconstrução de áreas ausentes em imagens faciais usando a técnica de Inpainting baseada em modelo de difusão*. [Master's thesis, Universidade Federal do Acre], UFAC.

Amit, T., Shaharbany, T., Nachmani, E., & Wolf, L. (2021). *Segdiff: Image segmentation with diffusion probabilistic models.* ArXiv. https://doi.org/10.48550/arXiv.2112.00390

Avrahami, O., Lischinski, D., & Fried, O. (2022). Blended diffusion for text-driven editing of natural images. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), New Orlean, LA, USA, 18208 – 18218. https://doi.org/10.48550/arXiv.2111.14818

Awais, M., Li, W., Hussain, S., Cheema, M. J. M., Li, W., Song, R., & Liu, C. (2022). Comparative evaluation of land surface temperature images from unmanned aerial vehicle and satellite observation for agricultural areas using in situ data. *Agriculture*, 12(2), 184. https://doi.org/10.3390/agriculture12020184

Ayala, C., Sesma, R., Aranda, C., & Galar, M. (2023). Diffusion models for remote sensing imagery semantic segmentation. *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium,* Pasadena, CA, USA (5654 – 5657). https://doi.org/10.1109/IGARSS52108.2023.10281461

Bandara, W. G. C., Nair, N. G., & Patel, V. M. (2022). *DDPM-CD: Denoising Diffusion Probabilistic Models as Feature Extractors for Change Detection.* ArXiv. https://doi.org/10.48550/arXiv.2206.11892

Bastiaanssen, W. G., Menenti, M., Feddes, R. A., & Holtslag, A. A. M. (1998). A remote sensing surface energy balance algorithm for land (SEBAL). 1. Formulation. *Journal of Hydrology*, *212*, 198-212. https://doi.org/10.1016/S0022-1694(98)00253-4

Bezerra, E., Mafalda, S., Alvarez, A. B., Uman-Flores, D. A., Perez-Torres, W. I., & Palomino-Quispe, F. (2023). A cloud coverage image reconstruction approach for remote sensing of temperature and vegetation in amazon rainforest. *Applied Sciences*, *13*(23), Article 12900. https://doi.org/10.3390/app132312900

Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P. S., & Sun, L. (2023). *A comprehensive survey of AI-Generated Content (AIGC): A history of generative AI from GAN to ChatGPT.* ArXiv. https://doi.org/10.48550/arXiv.2303.04226

Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, *34*, 8780-8794. https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf

Elharrouss, O., Almaadeed, N., Al-Maadeed, S., & Akbari, Y. (2020). Image inpainting: A review. *Neural Processing Letters*, *51*, 2007-2028. https://doi.org/10.1007/s11063-019-10163-0

Hidalgo García, D., & Arco Díaz, J. (2021). Spatial and multi-temporal analysis of land surface temperature through Landsat 8 images: comparison of algorithms

in a highly polluted city (Granada). *Remote Sensing*, *13*(5), 1012. https://doi.org/10.3390/rs13051012

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, & H. Lin (Eds.). *Advances in Neural Information Processing Systems*, *33 (NeurIPS 2020)* (pp. 6840-6851). https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

Jing, R., Duan, F., Lu, F., Zhang, M., & Zhao, W. (2023). Denoising diffusion probabilistic feature-based network for cloud removal in Sentinel-2 imagery. *Remote Sensing*, *15*(9), Article 2217. https://doi.org/10.3390/rs15092217

Kawar, B., Elad, M., Ermon, S., & Song, J. (2022). Denoising diffusion restoration models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.). *Advances in Neural Information Processing Systems*, *35 (NeurIPS 2020)*, 23593-23606. https://doi.org/10.48550/arXiv.2201.11793

Kortli, Y., Jridi, M., Al Falou, A., & Atri, M. (2020). Face recognition systems: A survey. *Sensors*, *20*(2), Article 342. https://doi.org/10.3390/s20020342

Leher, Q. O. (2024). *Inpainting com Modelos Generativos Probabilísticos de Difusão para a Reconstrução de áreas de Interesse em Imagens Satelitais*. [Bachelor's thesis, Universidade Federal do Acre]. UFAC.

Li, X., Ren, Y., Jin, X., Lan, C., Wang, X., Zeng, W., Wang, X. & Chen, Z. (2023). *Diffusion models for image restoration and enhancement. A comprehensive survey.* ArXiv. https://doi.org/10.48550/arXiv.2308.09388

Liu, J., Yuan, Z., Pan, Z., Fu, Y., Liu, L., & Lu, B. (2022). Diffusion model with detail complement for super-resolution of remote sensing. *Remote Sensing*, *14*(19), article 4834. https://doi.org/10.3390/rs14194834

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., & Van Gool, L. (2022). RePaint: Inpainting using denoising diffusion probabilistic models. 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* , New Orleans, LA, USA, (pp. 11461-11471). https://doi.org/10.48550/arXiv.2201.09865

Muzammal, N. M. Ranasinghe, K., Khan,S., Hayat,M., Khan, F. S., & Yang M.-H. (2021). Intriguing properties of vision transformers. *Adv. Neural Info. Process. Syst.*, *34*. https://doi.org/10.48550/arXiv.2105.10497

Murfitt, J., & Duguay, C. R. (2021). 50 years of lake ice research from active microwave remote sensing: Progress and prospects. *Remote Sensing of Environment*, *264*, Article 112616. https://doi.org/10.1016/j.rse.2021.112616

Naseer, M. M., Ranasinghe, K., Khan, S. H., Hayat, M., Khan, F. S., & Yang, M-H. (2021). Intriguing properties of vision transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S.

Liang, & J. Wortman Vaughan (Eds.). *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)* (pp. 23296-23308). https://proceedings.neurips.cc/paper_files/paper/2021/file/c404a5adbf90e09631678b13b05d9d7a-Paper.pdf

Perez-Torres, W. I., Uman-Flores, D. A., Quispe-Quispe, A. B., Palomino-Quispe, F., Bezerra, E., Leher, Q., Paixão, T. & Alvarez, A. B. (2024). Exploratory analysis using deep learning for water-body segmentation of Peru's high-mountain remote sensing images. *Sensors*, *24*(16), article 5177. https://doi.org/10.3390/s24165177

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. 2022 IEEE/CVF *Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, (10684-10695). https://doi.org/10.48550/arXiv.2112.10752

Singh, A., & Vyas, V. (2022). A review on remote sensing application in river ecosystem evaluation. *Spatial Information Research*, *30*(6), 759-772. https://doi.org/10.1007/s41324-022-00470-5

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. *Proceedings of Machine Learning Research, 37*, (2256-2265). https://proceedings.mlr.press/v37/sohl-dickstein15.html

Tan, H., Wu, S., & Pi, J. (2022). Semantic diffusion network for semantic segmentation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, & K. Cho and A. Oh (Eds.). *Advances in Neural Information Processing Systems 35 (NeurIPS 2020)* https://proceedings.neurips.cc/paper_files/paper/2022/file/396446770f5e8496ca1feb02079d4fb7-Paper-Conference.pdf

Taskiran, M., Kahraman, N., & Erdem, C. E. (2020). Face recognition: Past, present and future (a review). *Digital Signal Processing*, *106*, article 102809. https://doi.org/10.1016/j.dsp.2020.102809

Wu, J., Gan, W., Chen, Z., Wan, S., & Lin, H. (2023). *Ai-generated content (aigc): A survey.* ArXiv. https://doi.org/10.48550/arXiv.2304.06632

Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., & Yang, M.-H. (2023). Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, *56*(4), 1-39. https://doi.org/10.1145/3626235

Zhang, S., Li, J., & Yang, L. (2023). *Survey on controlable image synthesis with deep learning.* ArXiv. https://doi.org/10.48550/arXiv.2307.10275

Zhao, X., & Jia, K. (2023). Cloud removal in remote sensing using sequential-based diffusion models. *Remote Sensing*, *15*(11), Article 2861. https://doi.org/10.3390/rs15112861