Synthesizing SAR Images with Generative AI: Expanding to Large-Scale Imagery

Solène Debuysère, Nicolas Trouvé, Nathan Letheule, Elise Colin, Olivier Lévêque

Synthesizing SAR Images with Generative AI: Expanding to Large-Scale Imagery

1st Solène Debuysère
ONERA - DEMR
Université Paris Saclay
Palaiseau, France
solene.debuysere@onera.fr

2nd Nicolas Trouvé
ONERA - DEMR
Université Paris Saclay
Palaiseau, France
nicolas.trouve@onera.fr

3th Nathan Letheule
ONERA - DEMR
Université Paris Saclay
Palaiseau, France
nathan.letheule@onera.fr

4th Elise Colin
ONERA - DTIS
Université Paris Saclay
Palaiseau, France
elise.colin@onera.fr

5th Olivier Lévêque
ONERA - DEMR
Université Paris Saclay
Palaiseau, France
olivier.leveque@onera.fr

***Abstract*—This paper investigates the application of generative models in simulating radar images, highlighting their capacity in generating realistic images from text description. To overcome the usual limitation of image size constraint, we propose a model operating through a series of blocks to progressively enhance the image's resolution and details by leveraging a pre-trained latent diffusion model. This cascading architecture, structured into three levels, is augmented by integrating a latent scaling technique, enabling a gradual improvement in the scale of the image, thus allowing the generation of larger images. Our method also stands out for incorporating the ControlNet model, maintaining content consistency during the multi-resolution process. After training on aerial SAR images from the ONERA SETHI sensor, we compare our architecture with alternative methods for large-scale image generation using larger pre-trained models and outpainting methods. The results demonstrate a clear enhancement in the scale of simulated radar images while maintaining alignment with the contexts described by the text descriptions.**

***Index Terms*—Synthetic Aperture Radar (SAR), Generative Model, Diffusion Model, Earth Observation**

I. Introduction

Synthetic Aperture Radar (SAR) images provide invaluable all-weather acquisition capabilities across various operational scenarios, including environmental monitoring, change detection, and object recognition. However, their unconventional acquisition geometry, phenomenological behavior, and statistical characteristics pose challenges for analysis and interpretation. Furthermore, high-resolution SAR image acquisition, particularly in airborne configurations, remains costly and restricted, necessitating the simulation of synthetic images for effective algorithm assessment. In the SAR domain, generating realistic images poses a significant challenge due to the distinct physics involved in radar imagery compared to models pre-trained on optical images. Recent advancements in text-to-image models, such as Stable Diffusion (SD) [17], DALL-E-2 [6], Midjourney[12], and Imagen[18], offer the potential to produce synthetic images of relatively high quality and resolution. Yet, their full potential remains largely unexplored in terms of applications, computational efficiency, and evaluation metrics. Traditional generative models often encounter computational efficiency issues, particularly with large images and diffusion-based models in pixel space. Our study aims to explore computationally efficient methods for producing large scale high resolution SAR images using latent text-to-image models. We compare two methods utilizing pre trained model from [17] as a base model: a cascaded latent diffusion architecture, and out-painting process. A last method based on a large U-net architecture (SDXL) [14] attempting to generate larger images directly.

These approaches balance computational resources, image coherence, and resolution enhancement effectively. We create training datasets using fine-tuned image-to-text models for automated captionning applied to ONERA's airborne SAR images at various resolutions. The first methodology involves a cascaded latent diffusion model comprising three blocks, each contributing to process the images at different resolution. This approach incorporates prompts to guide the learning process, utilizing three distincts UNet trained at different resolutions, a latent upscaler, and a ControlNet [20] for conditioning. Comparative analysis of our three architecture models for high-resolution image enlargement reveals the higher maturity of the cascade architecture. Finally our work discuss the lack of dedicated comparative analysis in existing literature, aiming to investigate the ability of these methods to leverage descriptive captions and translate precise textual structuring into accurate spatial structuring in generated images.

After introducing the state of the art in image generative models, this paper explores the diffusion model and stable diffusion, and discusses high-resolution SAR generation, including objectives, dataset, and methodology. The methodology covers our cascaded latent diffusion architecture, Stable Diffusion XL approach, and the outpainting approach. Finally, we present our results, discuss the challenges of evaluation methods for SAR image generation, and conclude the paper.

II. Related works

Deep Generative Models Deep Generative models (DGMs) merge generative models with deep neural networks to handle various high-dimensional data types like time series, images, text, and audio. The objective of any generative model is to approximate the observed data distribution D, derived from a finite set of samples drawn from the underlying distribution p_{data}. During training, the model discerns underlying patterns and features, enabling it to generate new data instances by sampling from this learned distribution. Model

selection hinges on task complexity, dataset characteristics, and available computational resources. In computer vision, deep generative models focus on generating realistic images from scratch. Different approaches include image-to-image, image-to-text, and text-to-image models, differing in input data format and transformation process for producing the output.

Image-to-Image Approaches This approach transforms an input image into a new output image by learning the high-dimensional distribution of pixels constituting real images. Variational Autoencoders (VAEs) [5] encode data efficiently into compact representations, while Generative Adversarial Networks (GANs) [1] generate high-fidelity images through a min-max game. Studies in deep learning have explored SAR image generation. Liu et al. [10], Jones et al. [4], and Mason et al. [11] have highlighted the potential of GANs and recurrent auto-encoder network architectures in generating realistic SAR images.

Image-to-Text Approaches The image-to-text model comprehends visual content and context, generating coherent textual descriptions that elucidate the relationship between detected objects. Leading models include BLIP [8], BLIP2 [7], VLM [16], and LLAVA [9]. BLIP employs a Vision-Language Pre-training (VLP) approach and eliminates noisy samples to generate high-quality captions. Trouvé et al. [19] found that BLIP, pre-trained and fine-tuned on SAR imagery, enhances semantic richness in generated captions.

Text-to-Image Approaches This method creates images from textual descriptions, requiring a deep understanding of language semantics and visual representation. Key models include GANs and Diffusion Models (DMs). Stable Diffusion, in particular, conducts diffusion on a latent representation, producing diverse, high-quality samples [18]. While Stable Diffusion has not been extensively studied in the SAR domain, diffusion models like DDPM have improved SAR image quality [3].

III. Diffusion Model and Stable Diffusion

Recent progress in image generation owes much to diffusion models, especially Denoising Diffusion Probabilistic Models (DDPMs) [2]. DDPMs advance by iteratively introducing noise to an image and then learning to reverse this procedure. Initially, DDPM's forward process follows a Markov chain, sequentially adding noise to an image over multiple steps. This can be mathematically expressed as:

$$q(x_1, \ldots, x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{1}$$

and

$$q(x_t|x_{t-1}) = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon_t \tag{2}$$

where β_t represents the variance of the noise added at each step (called scheduler), and ϵ_t is the noise. Different schedulers for β_t, such as linear and cosine schedules, have been explored in the original paper [2]. The authors showed that the cosine schedule offers superior performance with a smoother degradation and less loss of information in the input image. As $T \to \infty$, x_T is nearly an isotropic Gaussian distribution. To express x_t through x_0, and simplify the training process without computing each distribution, two additional terms have been defined:

$$\bar{\alpha}_t = \prod_{i=0}^{t} \alpha_i \quad \text{and} \quad \alpha_t = 1 - \beta_t$$

$$x_t = \sqrt{\hat{\alpha}_t}x_0 + \sqrt{1-\hat{\alpha}_t}\epsilon \tag{3}$$

$\bar{\alpha}_t$ is a product of the variance terms that accumulate over the diffusion process up to step t

The reverse process aims to reconstruct the original image distribution $q(x_0)$ by learning the distribution $q(x_{t-1}|x_t)$ with a parameterized model. This is achieved using a neural network, a UNet architecture, which predicts the noise added at each step and reverses the process:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{4}$$

Indeed, during the forward process the noise added is Gaussian, so we search for p_θ to be Gaussian. Thus, the UNet needs to learn the mean and variance of the previous distribution at time step t. Therefore, the training process involves minimizing the difference between the actual noise and the noise predicted by the network, aiming to improve the fidelity of the generated images.

Compared to DDPM, Stable Diffusion operates in latent space rather than pixel space with a Variational Autoencoder (VAE), which encodes an image into a compact latent representation. The latent space allows for more efficient manipulation and generation of images during the forward and reverse process:

$$z = E(x), \quad \tilde{x} = D(z) \tag{5}$$

where E is the encoder, D is the decoder, and z is the latent representation of image x. Images x are compressed by a factor of 8 to obtain z.

Moreover, the text prompts - description of the scene we want to generate - are converted into embeddings via CLIP's text encoder [15], resulting in a matrix where each token (word) is a high-dimensional vector. The model is limited to 75 words plus 2 tokens for start and end, leading to 77 token embeddings $\tau_\theta(y)$.

Stable Diffusion [17] works entirely in the latent space and learn to predict the noise added to the latent image representation (during the forward process) and tries to denoise it (reverse process) with a UNet network conditonned on the timestep representation and text encoding $\tau_\theta(y)$. Each timestep t is converted into an embedding via a function such as the sinusoidal position encoding (often used in transformers).

During the training process, the weights of the UNet and any associated models (like the text encoder) are updated to minimize the difference between the predicted noise and the actual noise:

$$L_{LDM} := \mathbb{E}_{\epsilon(x), y, \epsilon \sim \mathcal{N}(0,1), t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right] \tag{6}$$

where ϵ_θ represents the predicted noise, and $\tau_\theta(y)$ denotes the text embeddings. and

$$z_t = \sqrt{\hat{\alpha}_t} z_0 + \sqrt{1 - \hat{\alpha}_t}\epsilon$$

In Stable Diffusion, the image generation conditioned by the text embeddings, is made possible by an attention mechanism within the UNet architecture. The mechanism computes attention scores that determine how much each part of the image's latent representation should focus on different parts of the text embedding. For instance, if the text mentions specific objects, colors, or actions, the attention mechanism can guide the model to emphasize these elements in the generated image.

During inference, the trained model starts by generating a noisy latent image representation from a normal distribution. It then iteratively refines this image representation through a series of steps, each time reducing the noise slightly. At each step, additional Gaussian noise may be added $\sigma_t z$, and a the predicted noise component is subtracted $\epsilon_\theta(x_t, t)$, combined with a variance term $\frac{1}{\sqrt{\alpha_t}}$. This iterative process gradually transforms the initial noise into a coherent image representation that aligns with the learned data distribution.

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \alpha_t}} \epsilon_\theta(x_t, t) \right) + \sigma_t z \qquad (7)$$

Note that the specific formula used for each step can vary depending on the sampling method employed. And after all the T inference steps, we obtain the representation within the latent space of the generated image x_0. The VAE Decoder produces the final image which is typically 512×512 pixels in size for stable diffusion SD1.5.

IV. High-resolution SAR Generation

A. Objectives

In this study, we explore methods for generating high-resolution SAR images using latent diffusion models. Acknowledging the pre trained model's default limitation of a 512x512 image output, insufficient for large scale and high-resolution applications, we investigate alternative approaches to leverage its potential without starting from scratch. Our approach begins with the creation of a labeled training sets, derived from ONERA's airborne images at square resolutions of 160cm, 80cm, and 40cm. These datasets are used for fine-tuning our image-to-text captioning models, to fine-tune Stable Diffusion base model with pairs of SAR-text samples. Fine-tuning is essential for aligning the distinctive characteristics of SAR imagery, including speckle and bright spots, with the generative model framework. Firstly, we implement a cascaded latent diffusion architecture, integrating multiple resolutions in a cascading manner to enhance image detail across various scales. Following this, we assessed the potential of the Stable Diffusion XL model, which generates larger-scale images through an enlarged UNet model. Finally we explored an out-painting strategy. The criteria for comparison include spatial coherence and computational viability determined through expert evaluations.

B. Dataset

A 15-year collection of X-band SAR images from southern France, obtained via a Falcon 20 aircraft with the SETHI system, was processed and segmented, resulting in over 25,000 image samples at resolutions of 40 cm, 80 cm, and 160 cm. Matching optical images were paired with SAR images to form datasets. The captioning model BLIP was fine-tuned with optical images from the RSICD and UCM databases across these resolutions. This adaptation enabled the generation of textual captions for optical-SAR image pairs, culminating in triplets of images and captions tailored for model fine-tuning at each resolution.

C. Methodology

1) Cascaded latent diffusion architecture: As described in Figure 1, our model uses a cascaded latent diffusion process to produce high-resolution images from text, progressively enhancing image resolution from low (160 cm at 512x512) through intermediate (80 cm at 1024x1024) to high (40 cm at 2048x2048) detail levels. The model is fine-tuned at each stage on SAR images and captions, ensuring detailed SAR image representation across different resolutions.

In our inference process, we generate SAR images through a stepwise architecture, starting with the generation of a base latent image vector from text prompts using the Stable Diffusion model fine-tuned on 160 cm resolution SAR images (SD160). This vector is then upscaled and transformed into a 512x512 image via a VAE Decoder, with the ControlNet model ensuring textual fidelity and contextual accuracy. Next, the Intermediate Refinement block, utilizing the Stable Diffusion model fine-tuned on 80 cm resolution (SD80) and an Upscaler, enhances the resolution to 2048x2048, adding finer details. Finally, the process concludes with the Stable Diffusion model fine-tuned on 40 cm resolution (SD40) and ControlNet, achieving the highest quality and resolution. This multi-resolution approach results in images that are highly detailed and accurately reflect their textual prompts, with ControlNet maintaining consistency during the multi-scale process.

For these generated examples in Figure 2, we used the following prompts: $prompt_{positive}$: 'city, buildings, roads', $prompt_{negative}$: 'blurry, illustration, forest'.

In Figure 3, the following prompts were used to generate the respective images : 'a beach that is crowded with tourists, the sea, and hotels in front of the beach' (Example 1), 'forest, river' (Example 2), 'A busy city with highways between closely spaced buildings, displaying a lively urban environment.' (Example 3).

2) Stable Diffusion XL approach: The Stable Diffusion XL (SDXL) model features a UNet backbone that is three times larger than the one in Stable Diffusion SD1.5, primarily due to additional attention blocks and an expanded cross-attention context. This expansion is facilitated by a second text encoder, enhancing SDXL's capability to produce larger, more detailed, and contextually accurate images based on the input text. The model was not fully fine tuned due to hardware limitations : a low rank training method was instead used.

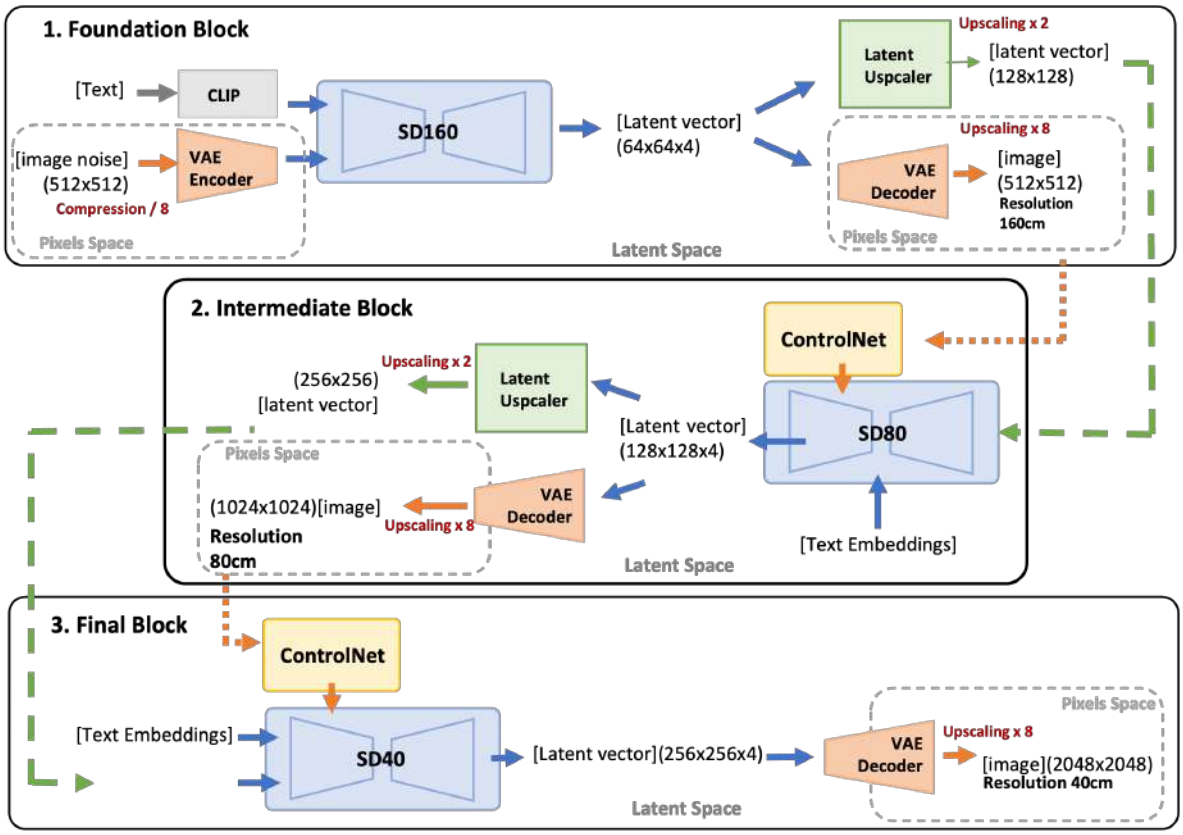


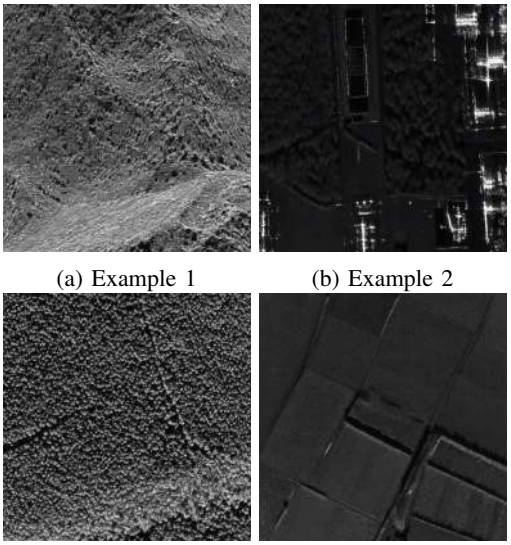
Fig. 1: Cascaded Latent Diffusion Architecture

(a) Block 512x512 (b) Block 1024x1024 (c) Block 2048x2048

Fig. 2: Model 1: Enhancing image resolution from low (160 cm at 512x512) through intermediate (80 cm at 1024x1024) to high (40 cm at 2048x2048) detail levels - Block generated outputs

(a) Example 1 (b) Example 2 (c) Example 3

Fig. 3: Model 1 - Different examples of generated images of size 2048x2048 generated with our cascade model using different prompts



(a) Example 1 (b) Example 2

(c) Example 3 (d) Example 4

Fig. 4: Model 2 - Different examples of generated images of size 2048x2048 generated with SDXL and LoRA model using different prompts

In the Figure 4, we used the following prompts to generate 2048x2048 images size with SDXL model : 'rugged mountain with sparse vegetation'(Example 1), 'large road surround by multiple buildings' (Example 2), 'a dense forest with sparse clearings' (Example 3), 'various fields with a small serpentine road' (Example 4).

While the results show the model ability to directly generate large realistic images, the texture (ie speckle) appears to be less faithful to the training material. Further works would involve training the full model to assess if those limitations are due to LoRa's method limitations, the base model itself, or the training set.

3) Outpainting approach: Outpainting, a variant of inpainting, involves editing a partially noised latent image to integrate it seamlessly with surrounding content according to a prompt, focusing on the expansion of an image's borders. This process, more complex due to its extrapolative nature, places the input image on a larger, artificially noised canvas to extend its dimensions. The study utilizes a basic U-net model, trained on fully noised images, for foundational inpainting, noting enhanced performance with training on images masked randomly. Examples demonstrate the method's ability to horizontally expand an image from 1024x1024 to 1024x2048 pixels effectively, though it falls short in overall coherence compared to other methods, suggesting the need for a specialized inpainting model to improve results.

In the provided examples in the Figure 5, an initial image measuring 1024x1024 pixels is horizontally expanded to create a 1024x2048 image. The conditional prompt used is 'city, buildings, roads'. These instances confirm the effectiveness of our method, highlighting its ability to expand images without introducing visible seams or artifacts at the junction of the original image and its extension. However, the overall coherence of the final image does not match the level achieved by other methods. Developing a specialized inpainting model would be required to further advance this concept.

D. Results

The first version of the pre-trained model from [17] faces challenges with model correlations and word placement in prompts, often limiting image diversity by linking specific attributes to certain themes. In contrast, SDXL performs better at generating large images without common distortions like 'double-heading,' handling larger dimensions more effectively. While SD1.5 can still produce detailed images, SDXL outperforms the model with the use of negative prompts and filters. Despite its superior composition, it still struggles with SAR texture quality, possibly due to the use of a compressed latent space and our fine-tuning approach with a single LoRA layer. This issue is exacerbated by a dataset size that is insufficient for the required scale of upscaling, needing significantly more data than provided. Additionally, while the outpainting diffusion method offers improved control over image expansion, it faces challenges in maintaining consistency and coherence at larger scales without training a dedicated model. More results, including examples of our three models' comparisons with different prompts and continuous seeds, are visible on the ONERA simulator Emprise website [13].

E. Challenges of evaluation methods

Evaluating the realism of Synthetic Aperture Radar (SAR) images generated by stable diffusion models presents unique challenges. Conventional evaluation metrics provide insights into the quality, diversity, and fidelity of generated images, but they often fall short in capturing the distinct characteristics of SAR data.

Score-based techniques like Precision and Recall and the Inception Score offer a preliminary assessment of image quality and variety. However, these methods do not adequately address the SAR-specific features that are crucial for realistic image generation. To evaluate the similarity between real and synthetic images, the Fréchet Inception Distance (FID) measures the distributional distance between features of real and generated images. Additionally, classifier-based evaluations can be trained to predict the realism of SAR images, focusing on complex noise patterns like speckle, which are inherent in SAR data. Further, some methods aim to evaluate the consistency and predictability of changes in generated images as we navigate the model's latent space. For example, the Perceptual Path Length (PPL) indicates a structured latent space if it is low, which is vital for ensuring that variations in generated images are coherent. However, these methods do not integrate the physical characteristics of SAR images into the evaluation process.

However, human judgments are still crucial, as statistical metrics alone cannot fully capture the complexity of realism perception. A survey mixing ONERA's SETHI data with images generated by our model can assess the perceptual judgment on realism from both experts and non-experts.

V. Conclusion

The use of latent representations in the latent diffusion architecture significantly boosts denoising and sampling processes, enhancing speed and robustness in image synthesis. Conditioning mechanisms like text prompts and ControlNet prove highly effective in guiding the generative process. The potential to condition on optical images or SAR imagery's physical properties opens new avenues for realism enhancement. Despite these successes, challenges persist in reproducing high-resolution SAR imagery. The gaussian distribution's mean parameter settings in forward diffusion models can disrupt white-level or very dark SAR images. Additionally, speckle poses a dual challenge, degrading image quality while contributing to SAR image realism, complicating its treatment in the generative process. Developing and applying metrics to evaluate simulated SAR image realism remains one of the main challenge.

In the field of high-resolution SAR image simulation using deep learning models, our exploration of three architectures — outpainting, Stable Diffusion XL (SDXL), and cascaded latent diffusion — reveals both promise and challenges. We assess compositional quality and texture fidelity, focusing on

(a) Real SAR Image (b) Outpainted image: example 1 (c) Outpainted image: example 2

Fig. 5: Initial image on the left, 1024 x 1024, followed by two outpainted versions, each corresponding to different generation seeds, resulting in images sized 1024 x 2048.

image coherence, adherence to textual prompts, and computational efficiency. Qualitative evaluation with SAR experts from ONERA indicates the higher maturity of the cascading architecture in enhancing resolution while maintaining coherence and fidelity. While SDXL enhances composition quality, texture fidelity requires refinement, suggesting the need for model improvement and dataset expansion. Outpainting, though versatile, requires careful context management to prevent loss. Our analysis underscores the importance of metrics in advancing generative models for SAR imagery synthesis, unlocking their potential for diverse applications in remote sensing and environmental monitoring.

REFERENCES

[1] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML].

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG].

[3] Xuran Hu et al. *SAR Despeckling via Regional Denoising Diffusion Probabilistic Model*. 2024. arXiv: 2401.03122 [cs.CV].

[4] Branndon Jones, Ali Ahmadibeni, and Amir Shirkhodaie. "Physics-based simulated SAR imagery generation of vehicles for deep learning applications". In: *Applications of Machine Learning 2020*. Vol. 11511. SPIE. 2020, pp. 162–173.

[5] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2022. arXiv: 1312.6114 [stat.ML].

[6] Sangyun Lee. "DALLE-2". In: ().

[7] Junnan Li et al. *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models*. 2023. arXiv: 2301.12597 [cs.CV].

[8] Junnan Li et al. *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. 2022. arXiv: 2201.12086 [cs.CV].

[9] Haotian Liu et al. *Improved Baselines with Visual Instruction Tuning*. 2023. arXiv: 2310.03744 [cs.CV].

[10] Wenlong Liu et al. "Generating simulated SAR images using generative adversarial network". In: *Applications of Digital Image Processing XLI*. Vol. 10752. SPIE. 2018, pp. 32–42.

[11] Eric Mason, Bariscan Yonel, and Birsen Yazıcı. "Deep learning for SAR image formation". In: *Defense + Security*. 2017. URL: https://api.semanticscholar.org/CorpusID:125941446.

[12] Midjourney. https://www.midjourney.com/home.

[13] DEMR-SEM ONERA. *Examples of AI Generated SAR Images*. Accessed: 2024-07-24. 2024. URL: https://www.emprise-em.fr/index.php/ai-generated-images/.

[14] Dustin Podell et al. *SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis*. 2023. arXiv: 2307.01952 [cs.CV].

[15] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV].

[16] Wes Robbins. "Towards Multimodal Vision-Language Models Generating Non-Generic Text". In: *ICON*. 2022. URL: https://api.semanticscholar.org/CorpusID:250299115.

[17] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. arXiv: 2112.10752 [cs.CV].

[18] Chitwan Saharia et al. *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*. 2022. arXiv: 2205.11487 [cs.CV].

[19] Nicolas Trouve et al. "SAR image synthesis using text conditioned pre-trained generative AI models". In: *Proceedings of EUSAR 2024; 15th European Conference on Synthetic Aperture Radar*. VDE. Munich, Germany: VDE,ITG, 2024.

[20] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. *Adding Conditional Control to Text-to-Image Diffusion Models*. 2023. arXiv: 2302.05543 [cs.CV].