Exploring the Capability of Text-to-Image Diffusion Models with Structural Edge Guidance for Multi-Spectral Satellite Image Inpainting

Mikolaj Czerkawski, Christos Tachtatzis
Department of Electronic and Electrical Engineering
University of Strathclyde, Glasgow, UK

***Abstract*—The paper investigates the utility of text-to-image inpainting models for satellite image data. Two technical challenges of injecting structural guiding signals into the generative process as well as translating the inpainted RGB pixels to a wider set of MSI bands are addressed by introducing a novel inpainting framework based on StableDiffusion and ControlNet as well as a novel method for RGB-to-MSI translation. The results on a wider set of data suggest that the inpainting synthesized via StableDiffusion suffers from undesired artefacts and that a simple alternative of self-supervised internal inpainting achieves higher quality of synthesis.**

***Index Terms*—image inpainting, image completion, generative models**

I. Introduction

In many circumstances, it may be required to inpaint regions of optical satellite images, due to common issues such as a sensor failure or natural conditions like cloud cover. This challenge has been widely addressed in the literature [1], [2], [3], [4], [5], however, the use of general pre-trained text-to-image diffusion models to solve this problem has not yet been explored.

It is not currently clear how beneficial these general-purpose text-to-image generative models, such as StableDiffusion [6], could be for the task of satellite image inpainting. Trained on a diverse set of data, containing both text and images, and optimized to solve the challenging task of text-to-image synthesis, they could be good candidates for sourcing meaningful inductive bias from other tasks. Furthermore, denoising diffusion models have several promising properties of their own, most prominently, the flexibility to trade-off compute cost and synthesis quality at inference time [7]. Furthermore, StableDiffusion [6] is compatible with ControlNet [8], which permits injection of a structural guidance image into the synthesis process, which can be used to condition the process on historical satellite data.

However, since the majority of text-to-image generators are designed for RGB data, the application to multi-spectral satellite image inpainting involves two challenges:

- Inpainting of the RGB channel subset using general-purpose text-to-image models
- Translation of the RGB inpainting into all multi-spectral bands

Furthermore, as mentioned earlier, it is often beneficial to guide the structure of the synthesized image using additional signals, such as historical samples. Hence, there is one more, albeit optional, challenge addressed here:

- Injection of structural guide in to the inpainting process

In this manuscript, two solutions are proposed and tested using off-the-shelf open-source text-to-image models to inpaint large portions of multi-spectral 13-band Sentinel-2 images, by employing a two-stage approach consisting of (1) RGB-based inpainting and (2) RGB-to-MSI zero-shot translation, as illustrated in Figure 1. A set of experiments with various hyperparameter settings is provided to gain an understanding of the influence these have on the proposed framework. This is followed by an evaluation of the proposed inpainting methods on multi-spectral satellite image data.

II. Related Work

This work considers methods that can be freely applied to multi-spectral image representations and adapt to their content and channel count (assuming their representation includes the RGB channels). Existing pre-trained models, such as the one based on convolutional neural processes for inpainting [10], do not provide such a level of freedom and hence remain out of the scope of this work.

Denoising Diffusion Models. The approach to use denoising diffusion processes for image generation has become popular over the last few years [11], [12], [13], [14], [15]. While more expensive at inference, diffusion models have been shown to beat the state of the art in terms of image quality [15]. Other advantages include control over the trade-off between quality and inference time, as well as ways of injecting many types of guidance to condition the synthesis [15]. Eventually, text prompts became one of the most common ways of conditioning, giving rise to text-to-image models such as DALLE-2 [16], Imagen [17], or StableDiffusion [6]. The latter, StableDiffusion, was the first open-source implementation of such a model with free and open access to model parameters, which this work benefits from.

Single Image Synthesis for Satellite Images. Solutions relying on learning based on the input sample only can often be applied to any type of data modality and shape. For the task of satellite image inpainting, several works have explored this direction [3], [18] by relying on priors captured by the a convolutional neural network topology, along the lines of the seminal work on Deep Image Prior [9]. However, none of

This is a peer-reviewed, version of record of the following paper: Czerkawski, M., & Tachtatzis, C. (2024). Exploring the capability of text-to-image diffusion models with structural edge guidance for multi-spectral satellite image inpainting. IEEE Geoscience and Remote Sensing Letters. Advance online publication. https://doi.org/10.1109/LGRS.2024.3370212

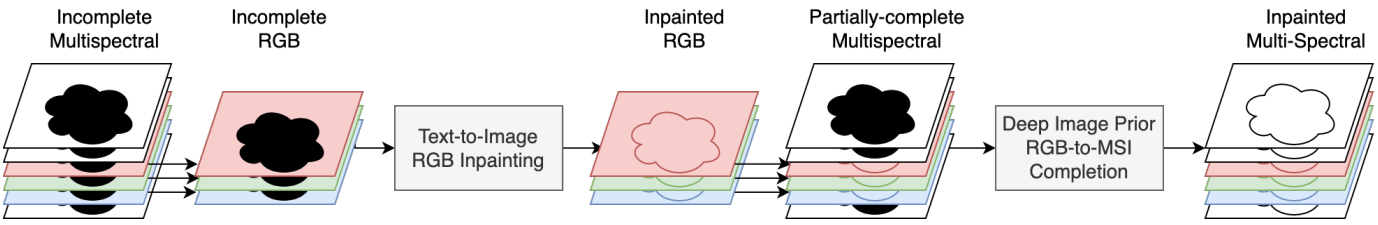

Fig. 1. Complete pipeline for multi-spectral satellite image inpainting. The process is built on a sequence of two steps, where a pre-trained diffusion model is first applied to RGB data for inpainting, and then a Deep Image Prior [9] approach is used to transfer that data into non-RGB channels.

the past approaches has explored the potential of combining single image synthesis with the general-purpose text-to-image diffusion models.

III. Method

A. Components

StableDiffusion is a variant of latent diffusion [6] focused on the text conditioning modality. Latent diffusion is a special type of image diffusion aimed at high-resolution data, which employs an autoencoder to compress image input into a latent space so that the denoising diffusion process is performed in a more compact domain than high-resolution images. StableDiffusion refers not only to the method (which is technically the same as latent diffusion) but also to the model weights released by the StabilityAI organisation. So far, several models have been released, with different levels of performance, but also, as in the case of StableDiffusion-2, a different set of potential inputs; for example, depth image condition. In this work, the StableDiffusion1.5 checkpoint is used with the model variant trained to perform inpainting based on an additional inpainting mask provided in input.

Diffusion for Inpainting. Several approaches for employing denoising diffusion for the task of image inpainting have already been considered. Some prominent examples include Palette [17], where the network performing denoising diffusion in pixel space is also provided with an additional channel corresponding to the inpainting mask, effectively serving as an extra condition present in the input. An alternative approach of RePaint [19] uses a pre-trained denoising diffusion model for inpainting with a different type of sampling scenario, where the known regions of the input (passed through the forward process) are mixed with the diffused signal representation. Finally, StableDiffusion also provides inpainting-oriented variants, where the underlying latent space model accepts an extra channel for the downsampled mask besides the usual 4-channel latent representation [6]. As mentioned earlier, the same approach was used for the inpainting model of StableDiffusion, the inpainting core model for this work.

ControlNet. Another important component of the presented approach is ControlNet, which enables to inject additional spatial guidance into the inpainting process. ControlNet has been introduced [8] an extension to StableDiffusion with the aim of incorporating more image-based conditions into the synthesis process. More specifically, the method uses a separate encoding network to encode latents based on a preselected condition type and mixes the encoded representations with the internal representations in the core StableDiffusion model. To ensure preservation of the features learned by the core StableDiffusion model, a zero-convolution technique has been used to nullify the effect of the added network at the beginning of the fine-tuning process. In the seminal ControlNet paper [8], several choices for the conditioning signal type are proposed, including Canny edge, Hough lines, user sketches, human pose, or HED boundary detections [20].

B. Edge-Guided Inpainting for RGB.

This work combines the inpainting capability of the StableDiffusion model with the edge-guidance capability of a HED boundary version of ControlNet, named **Edge-Guided Inpainting**. As shown in Figure 2, this is achieved by combining a pre-trained ControlNet model with the StableDiffusion inpainting backbone. This allows synthesis of the inpainted area, based on the provided mask input, control image and an optional text-prompt. This work focuses primarily on the use case with a historical edge-guidance, and a simple prompt of ”a cloud-free satellite image”. However, the proposed framework provides a large degree of flexibility between choosing different text-prompts (including negative prompts) to guide the output as well as different sources of structural information other than historical image edges.

C. RGB-to-MSI transfer with Deep Image Prior

The problem of generating multi-spectral images from an RGB sample has been explored in earlier works such as [21], [22], [23], [24], however, these solutions generally require training a specialised network for every image modality, and are not well adjusted for this case, where part of the multi-spectral image is known. To enable a flexible, representation-agnostic approach, a Deep Image Prior [9] technique is used, where a randomly initialized convolutional network is optimized to produce the available pixels in the output, similar to previous work on satellite image inpainting [18]. A SkipNetwork with the same architecture as in [18] is optimized with MSE loss backpropagated from the known region for 4,000 gradient steps at a learning rate of 0.02. The known region contains the pixels of all bands that are not missing as well as a complete inpainted image in the RGB bands, arranged in the same fashion as the input representation.

IV. Results

The test dataset used for this study is extracted from SEN12MS-CR-TS [25] containing 888 cloud-free Sentinel-2

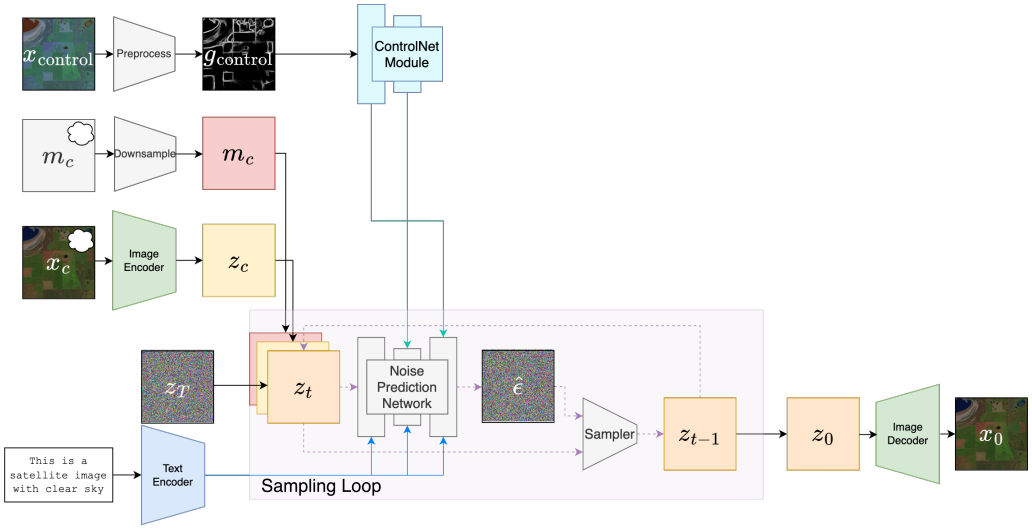

Fig. 2. The **Edge-Guided Inpainting** diffusion pipeline used for this work employs a ControlNet approach [8], with an inpainting StableDiffusion backbone.

TABLE I
PARAMETER TEST RESULTS FOR THE TEXT-BASED MODELS.

		StableDiffusion Inpainting				Edge-Guided Inpainting			
		SSIM (↑)		RMSE (↓)		SSIM (↑)		RMSE (↓)	
		Whole	Mask	Whole	Mask	Whole	Mask	Whole	Mask
Mask Content	Blank	0.70	0.54	0.11	0.15	0.71	0.58	0.10	0.14
	Historical	0.78	0.67	0.10	0.13	0.77	0.67	0.10	0.13
Text-Guidance Scale	0.0	0.78	0.67	0.10	0.13	0.77	0.67	0.10	0.13
	1.0	0.78	0.67	0.10	0.13	0.77	0.67	0.10	0.13
	7.5	0.77	0.66	0.11	0.14	0.78	0.68	0.09	0.12
Sampling Steps	**20**	0.78	0.67	0.10	0.13	0.77	0.67	0.10	0.13
	50	0.77	0.66	0.10	0.13	0.77	0.66	0.10	0.13
	100	0.77	0.66	0.10	0.13	0.76	0.66	0.10	0.13
Edge-Guidance Scale	0.1	NA				0.78	0.67	0.10	0.13
	0.5	NA				0.79	0.69	0.09	0.12
	1.0	NA				0.77	0.67	0.10	0.13

test samples, each paired with another historical cloud-free Sentinel-2 sample from the exact same location, which is used as the historical structure guide signal. The raw data is subject to the same preprocessing as in the related dataset [2], followed by a clipping operation to constrain the samples to [0,1] range expected by StableDiffusion. It was also ensured that the mean value of the samples is not higher than 0.9 before the clipping operation to exclude saturated images from analysis.

The two tested text-based models include the standard StableDiffusion inpainting approach and the proposed Edge-Guided Inpainting approach.

A. Parameter Tests

The following parameters of the text-to-image model are tested: 1) the content of the masked region 2) the text-guidance scale 3) the number of sampling steps 4) the edge-guidance scale (which only applies to the edge-guided inpainting approach). The results are shown in Table I. For each tested parameter, the remaining parameters have the values highlighted in bold.

Since this study is focused on the utility of historical optical data, explored is an approach to fill the missing regions with the values extracted from the historical sample instead of leaving them blank, in order to inject structure information into the network input. An example of these two input variants is shown in Figure 3.

It is found that the historical input injection is beneficial and can improve the output structure, as indicated by the increased in SSIM in Table I. This effect occurs standard StableDiffusion inpainting (inpainted SSIM goes from 0.54 to 0.67) and Edge-Guided Inpainting (inpainted SSIM goes from 0.58 to 0.67). Without using this technique Edge-Guided inpainting performs better at reconstructing the missing region, owing to the additional guidance from the historical sample using ControlNet. However, the technique of historical input filling puts both standard and ControlNet inpainting variants on par.

The classifier-free guidance scale for the text prompt, as defined in [15], controls the influence of the text input. and has been tested at three distinct levels, ranging from 0.0 (no effective text guidance) through 1.0 to 7.5 (StableDiffusion default). Since additional information is supplied in the form of historical visual data (either via ControlNet or input historical filling), the differences are very small. The text appears to be less useful for the standard StableDiffusion inpainting, with the performance slightly higher for low levels of text-guidance

Fig. 3. Comparison of the two methods of filling the masked region in the input to the diffusion models. Furthermore, output achieved with the StableDiffusion Inpainting scheme is shown for reference as a result of using each method.

scale, while for the Edge-Guided Inpainting model it is the opposite.

The UniPC sampler [26] is used for all diffusion sampling and the main tested factor is the number of sampling steps, with tested values of 20 steps (reported in [26] to yield good quality), 50 steps and 100 steps, verifying that the increased number of steps is not found to be beneficial for this task.

The weight applied to the ControlNet features [8] before they are added to the core network features is tested with the default value of 1.0, along with 0.5 and 0.1 values that explore a more subtle conditioning scheme. It is found that from the tested values, 0.5 achieves the highest output quality and is used for the subsequent experimentation.

TABLE II
INPAINTING RESULTS COMPUTED **FOR ALL 13 CHANNELS** OF THE MULTISPECTRAL IMAGES IN THE TEST DATASET.

Method	SSIM (↑)		RMSE (↓)	
	Whole	Mask	Whole	Mask
SD-Inpainting	0.78	0.65	0.16	0.21
Edge-Guided Inpainting	0.62	0.48	0.37	0.48
Direct-DIP	0.64	0.45	0.38	0.53
Direct-DIP w/ Historical	0.85	0.74	0.14	0.19
Ideal-RGB Channel Fill	0.89	0.82	0.12	0.16

TABLE III
INPAINTING RESULTS COMPUTED **ONLY FOR THE RGB CHANNELS** OF THE MULTISPECTRAL IMAGES IN THE TEST DATASET.

Method	SSIM (↑)		RMSE (↓)	
	Whole	Mask	Whole	Mask
SD-Inpainting	0.78	0.67	0.10	0.13
Edge-Guided Inpainting	0.79	0.69	0.09	0.12
Direct-DIP	0.72	0.58	0.23	0.31
Direct-DIP w/ Historical	0.88	0.79	0.08	0.11

B. Multi-Spectral Inpainting Evaluation

The two proposed two-stage approaches based on text-to-image inpainting are compared against two single-stage methods applying direct inpainting with Deep Image Prior (similar to [18]), with the performance on multi-spectral images listed in Table II. In the last row, the RGB-to-MSI method is also tested (Ideal-RGB) with exact knowledge of RGB channels to approximate the maximum potential performance achievable with knowledge of the RGB channels, with 0.82 mask SSIM achieved. The best method is the Direct-DIP approach supplemented with historical data, with a SSIM of 0.74, followed by the text-to-image methods scoring 0.65 and 0.48, respectively.

The advantage of the DIP-based benchmark is also observed for the RGB channels alone, as shown in Table III.

The visual examples shown in Figure 4 illustrate the behavior of each framework, showing that for large inpainting masks, the text-to-image methods (last two rows) tend to introduce a lot of new objects into the scene, despite the structural guidance. For smaller masks, they appear to generate more convincing generations than DIP-based approaches.

V. CONCLUSIONS

Even with structure-oriented adjustments, the general-purpose text-to-image diffusion models may not be immediately performant for the task of multi-spectral satellite image inpainting. They tend to produce visually pleasing images, but their synthetic capability is prone to generate unnatural artefacts in the output. At this point, even a simple baseline of internal inpainting with historical data yields higher performance.

Despite this limitation, the application of text-based models may still be deemed attractive and motivate further work on the topic. It offers a convenient conditioning mechanism with text prompts, which could enable a variety of different applications, including a more controlled restoration process, or data augmentation.

REFERENCES

[1] P. Singh and N. Komodakis, "Cloud-GAN: Cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks," *International Geoscience and Remote Sensing Symposium (IGARSS)*, vol. 2018-July, pp. 1772–1775, 2018.

[2] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in Sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, no. January, pp. 333–346, 2020. [Online]. Available: https://doi.org/10.1016/j.isprsjprs.2020.05.013

[3] P. Ebel, M. Schmitt, and X. X. Zhu, "Internal Learning for Sequence-to-Sequence Cloud Removal via Synthetic Aperture Radar Prior Information," *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pp. 2691–2694, 2021.

[4] Z. Xu, K. Wu, W. Wang, X. Lyu, and P. Ren, "Semi-supervised thin cloud removal with mutually beneficial guides," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 192, no. August, pp. 327–343, 2022. [Online]. Available: https://doi.org/10.1016/j.isprsjprs.2022.08.026

[5] A. Sebastianelli, E. Puglisi, M. P. D. Rosso, J. Mifdal, A. Nowakowski, P. P. Mathieu, F. Pirri, and S. L. Ullo, "PLFM: Pixel-Level Merging of Intermediate Feature Maps by disentangling and fusing spatial and temporal data for Cloud Removal," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–1, 2022.

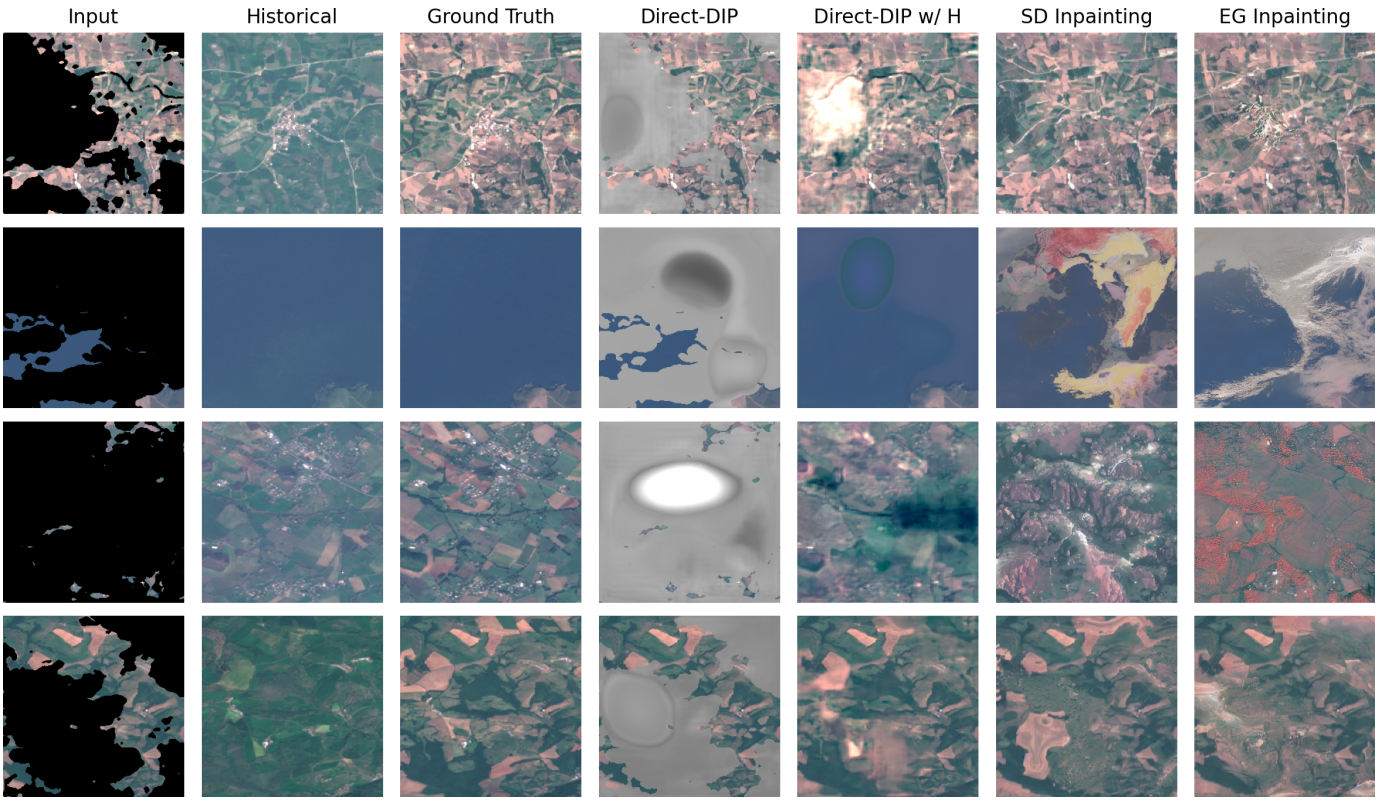

Fig. 4. RGB visualization of 4 random samples drawn from the test dataset and the corresponding output from each method. It is shown that the Direct-DIP struggles to perform good quality inpainting with no extra source of information, producing visually incoherent output. The text-based models appear to produce visually coherent, yet inaccurate inpaintings, despite the efforts to inject correct structural information into the process.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June, pp. 10 674–10 685, 2022. [Online]. Available: http://arxiv.org/abs/2112.10752

[7] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," *International Conference on Machine Learning*, 2023.

[8] L. Zhang and M. Agrawala, "Adding Conditional Control to Text-to-Image Diffusion Models," 2023. [Online]. Available: http://arxiv.org/abs/2302.05543

[9] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep Image Prior," *International Journal of Computer Vision*, vol. 128, no. 7, pp. 1867–1888, 2020.

[10] A. Pondaven, M. Bakler, D. Guo, H. Hashim, M. Ignatov, and H. Zhu, "Convolutional Neural Processes for Inpainting Satellite Images," no. 2015, 2022. [Online]. Available: http://arxiv.org/abs/2205.12407

[11] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *32nd International Conference on Machine Learning, ICML 2015*, vol. 3, 2015, pp. 2246–2255.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 2020-December, no. NeurIPS 2020, 2020, pp. 1–25.

[13] J. Song, C. Meng, and S. Ermon, "Denoising Diffusion Implicit Models," pp. 1–22, 2020. [Online]. Available: http://arxiv.org/abs/2010.02502

[14] A. Nichol and P. Dhariwal, "Improved Denoising Diffusion Probabilistic Models," 2021.

[15] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," *Advances in Neural Information Processing Systems*, vol. 11, pp. 8780–8794, 2021.

[16] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," no. Figure 3, 2022. [Online]. Available: http://arxiv.org/abs/2204.06125

[17] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi, "Palette: Image-to-Image Diffusion Models," *Proceedings of ACM SIGGRAPH*, vol. 1, no. 1, pp. 1–10, 2022. [Online]. Available: http://arxiv.org/abs/2111.05826

[18] M. Czerkawski, P. Upadhyay, C. Davison, A. Werkmeister, J. Cardona, R. Atkinson, C. Michie, I. Andonovic, M. Macdonald, and C. Tachtatzis, "Deep Internal Learning for Inpainting of Cloud-Affected Regions in Satellite Imagery," *Remote Sensing*, vol. 14, no. 6, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/6/1342

[19] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "RePaint: Inpainting using Denoising Diffusion Probabilistic Models," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June, pp. 11 451–11 461, 2022. [Online]. Available: http://arxiv.org/abs/2201.09865

[20] S. Xie and Z. Tu, "Holistically-Nested Edge Detection," *International Journal of Computer Vision*, vol. 125, no. 1-3, pp. 3–18, 2017.

[21] R. M. Nguyen, D. K. Prasad, and M. S. Brown, "Training-based spectral reconstruction from a single RGB image," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8695 LNCS, no. PART 7, pp. 186–201, 2014.

[22] J. Wu, J. Aeschbacher, and R. Timofte, "In Defense of Shallow Learned Spectral Reconstruction from RGB Images," *Proceedings - 2017 IEEE International Conference on Computer Vision Workshops, ICCVW 2017*, vol. 2018-January, pp. 471–479, 2017.

[23] X. H. Han, B. Shi, and Y. Zheng, "Residual HSRCNN: Residual Hyper-Spectral Reconstruction CNN from an RGB Image," *Proceedings - International Conference on Pattern Recognition*, vol. 2018-August, pp. 2664–2669, 2018.

[24] T. Zeng, C. Diao, and D. Lu, "U-Net-Based Multispectral Image Generation from an RGB Image," *IEEE Access*, vol. 9, no. 2, pp. 43 387–43 396, 2021.

[25] P. Ebel, Y. Xu, M. Schmitt, and X. X. Zhu, "SEN12MS-CR-TS: A Remote Sensing Data Set for Multi-modal Multi-temporal Cloud Removal," *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

[26] W. Zhao, L. Bai, Y. Rao, J. Zhou, and J. Lu, "UniPC: A Unified Predictor-Corrector Framework for Fast Sampling of Diffusion Models," 2023. [Online]. Available: http://arxiv.org/abs/2302.04867