Text-Prompted Diffusion Null-Space Model for Zero-Shot Image Restoration

Haijing ZHANG, Yining MAO

Abstract—In digital image processing, restoring a high-quality image from its degraded version is a longstanding challenge. Traditional methods often require either detailed knowledge of the degradation process or extensive datasets tailored for each specific restoration task. With the advent of generative models like Denoising Diffusion Probabilistic Models (DDPMs), new paths have emerged for image restoration. However, these innovative models usually need task-specific adjustments, which can limit their wider application.
This project explores a novel concept: combining the Denoising Diffusion Null-Space Model (DDNM) with text-prompted latent diffusion models to enhance zero-shot image restoration. By utilizing the capabilities of latent diffusion models and the direction provided by text prompts, we aim to refine the outcomes of diffusion-based restoration methods, which sometimes yield less-than-ideal results due to their stochastic nature. Through experiments in colorization, inpainting, and deblurring, our approach shows promise, indicating compatibility with latent space models and a potential improvement in the quality of restored images. This endeavor represents a step toward more flexible and effective solutions for image restoration, highlighting the importance of ongoing research and experimentation in the field.ke adjustments to the restoration results.

Index Terms—Computational Photography, Image Restoration, Diffusion Model, Stable Diffusion

1 INTRODUCTION

In the realm of digital image processing, image restoration remains a cornerstone challenge, addressing the recovery of an original image from a degraded version. The degradation can stem from various sources, including noise, blur, compression artifacts, and missing pixels, commonly arising during image acquisition, transmission, and storage. Traditional image restoration techniques have relied heavily on model-based and supervised learning approaches, which, while effective, often require specific knowledge about the degradation process or extensive labeled datasets for each task [1].

Recent advancements in deep learning have ushered in a new era for image restoration, particularly through the development of generative models. Among these, Denoising Diffusion Probabilistic Models (DDPMs) have shown promising results, simulating the diffusion process to generate high-quality images from noise [2]. However, these models typically necessitate task-specific adaptations or training regimes, limiting their flexibility and generalizability.

Emerging from this backdrop is the Zero-Shot Image Restoration (ZSIR) paradigm, which aims to leverage pre-trained models to restore images without task-specific training. A noteworthy contribution to this field is the Denoising Diffusion Null-Space Model (DDNM), which proposes a zero-shot framework for arbitrary linear image restoration tasks using a pre-trained diffusion model [3]. DDNM distinguishes itself by refining only the null-space contents during the reverse diffusion process, allowing for data-consistent and realistic restoration outcomes.

• *Haijing ZHANG and Yining Mao are with the Department of Electrical Engineering, Stanford University, Stanford, CA, 30332.*
E-mail: haijing@stanford.edu; yiningm@stanford.edu

Building on the foundation laid by DDNM and the burgeoning potential of latent diffusion models for high-resolution image synthesis, our project introduces a novel approach: integrating the Denoising Diffusion Null-Space Model with text-prompt-enabled latent diffusion models, specifically Stable Diffusion [4]. This integration aims to harness the descriptive power of text prompts to guide the restoration process, addressing the inherent randomness in diffusion models and their occasional generation of undesirable results.

This project report delineates our exploration into enhancing the zero-shot image restoration capabilities of DDNM through the incorporation of text prompts. By re-implementing both the DDNM and latent diffusion algorithms and conducting experiments across various restoration tasks—namely colorization, inpainting, and deblurring—we endeavor to demonstrate the feasibility and efficacy of our approach. Our findings suggest that this innovative integration not only maintains compatibility with latent space models but also significantly improves the flexibility and robustness of image restoration outcomes.

In the following sections, we first review the related work to set the stage for our contributions by examining existing techniques in image restoration. We then introduce our theory/method, detailing our approach that integrates the Denoising Diffusion Null-Space Model with text-prompted latent diffusion models for zero-shot image restoration. Our analysis & evaluation segment not only explores the methodology behind our experiments but also compares our results with those of other methods, highlighting the advantages and novelties of our approach. The results section presents both qualitative and quantitative outcomes, demonstrating the effectiveness of our method across various tasks. Finally, in the discussion section, we

Fig. 1. Our generated results on Image Restoration tasks using Stable Diffusion model and DDNM method. From left to right: Colorization, Inpainting and Deblurring.

reflect on the limitations of our current work, speculate on future work that could extend our findings, and conclude by summarizing the significance of our contributions to the field of image restoration.

2 Related Work

2.1 Image Restoration.

Aiming to recover a high-quality image from its degraded observation, image restoration (IR) tasks such as super-resolution (SR), deblurring, denoising, inpainting, and artifact removal have been long-standing research topics. Over the years, image restoration techniques have undergone significant evolution, transitioning from methods that rely on predefined degradation models, like bicubic downsampling, to more sophisticated approaches that tackle real-world complexities.

The first method involves model-based techniques that typically formulate the problem as an optimization task. These methods leverage a combination of data fidelity and regularization terms, with the latter incorporating prior knowledge about natural image statistics. While these approaches are interpretable and can often be solved efficiently, they may not fully capture the complexity of real-world image distributions, leading to results that lack high-frequency details. The mathematical formulation can be represented as:

$$\hat{x} = \arg\min_{x} \left(\frac{1}{2\sigma^2} ||Ax - y||_2^2 + \lambda R(x) \right) \quad (1)$$

This equation represents the trade-off between fidelity to the degraded image y through the term $||Ax - y||_2^2$ and the regularization $R(x)$ that encodes prior knowledge about the image distribution.

Subsequently, deep learning-based methods, especially end-to-end trained convolutional neural networks (CNNs), have shown remarkable success in various IR tasks. By directly learning the mapping from degraded to high-quality images from large datasets, these methods have surpassed traditional techniques in performance. However, they tend to be data and task-specific, requiring retraining for each new IR task or degradation type, which limits their flexibility and generalizability. The deep learning approach is typically formulated as:

$$\hat{x} = \arg\min_{\theta} \sum_{i=1}^{N} ||D_\theta(y_i) - x_i||_2^2 \quad (2)$$

Here, D_θ represents the deep network parameterized by θ, which is trained to minimize the difference between the network's output and the ground truth high-quality image x.

More recently, methods leveraging the latent spaces of pretrained generative models have emerged. These approaches, which can be seen as zero-shot in nature, optimize the latent codes of generative models to reconstruct high-quality images that are both consistent with the observed degradation and realistic. While offering a more flexible and task-agnostic solution, balancing the trade-off between fidelity to the degraded input and the realism of the output remains a challenging aspect, occasionally resulting in suboptimal restorations. This is mathematically formulated as:

$$\hat{x} = \arg\min_{w} \left(\frac{1}{2\sigma^2} ||AG(w) - y||_2^2 + \lambda R(w) \right) \quad (3)$$

In this formulation, $G(w)$ represents the generative model's output for the latent code w, and the objective is to find the w that minimizes the difference between the generated image $AG(w)$ and the degraded observation y, while also ensuring that w leads to a realistic image.

Each of these methodologies represents a unique approach to addressing the intricacies of image restoration, highlighting the field's dynamic evolution as it seeks to bridge the gap between theoretical models and real-world application requirements.

2.2 Deep-learning Models for Image Restoration

The advent of deep learning has revolutionized the field of image restoration with the introduction of sophisticated architectures, particularly Convolutional Neural Networks (CNNs) [5] and Transformers [6]. These architectures have been meticulously crafted to tackle complex degradation patterns. Moreover, the emergence of generative adversarial networks (GANs) [7] has marked a significant milestone in the realm of unsupervised learning for image restoration. GANs excel in synthesizing realistic Low-Resolution to High-Resolution (LR-HR) image pairs, adeptly addressing the nuances of blind super-resolution problems by generating explicit degradation models. This paradigm shift towards utilizing deep-learning models emphasizes the strategic design of model architectures and their training methodologies to enhance image restoration outcomes.

2.3 Diffusion-model Based Image Restoration

In the domain of image generation and restoration, diffusion models [8] have showcased remarkable capabilities. These models excel in image restoration by employing a gradual denoising process, which is inherently suited for reconstructing high-quality images from degraded inputs. The foundation of diffusion models lies in their unique structure and algorithmic principles, which facilitate an iterative refinement process. This process is crucial for achieving high-fidelity image synthesis and restoration. Notably, the integration of diffusion models into image super-resolution, as demonstrated in StableSR [9] using the Stable Diffusion model [10], highlights their transformative potential in the image restoration landscape. Further expanding the utility of diffusion models, the Denoising Diffusion Null-Space Model (DDNM) [11] introduces a novel zero-shot approach to image restoration. DDNM leverages the inherent strengths of diffusion models for generating realistic images without the need for task-specific training. By refining only the null-space content during the reverse diffusion process, DDNM offers a flexible and robust solution for diverse image restoration tasks, underscoring the benefits of zero-shot learning and the adaptability of diffusion models in addressing complex degradation scenarios.

3 Method

This project proposes an approach to image restoration by leveraging the capabilities of the Denoising Diffusion Null-Space Model (DDNM) and integrating it with Stable Diffusion, guided by textual prompts. Our method extends traditional image restoration techniques, allowing for flexible and robust zero-shot restoration across various tasks such as colorization, inpainting, and deblurring. Below, we detail the foundational components and the integrated pipeline of our method.

3.1 Denoising Diffusion Null-Space Model (DDNM)

DDNM serves as the cornerstone of our method, introducing a zero-shot framework for arbitrary linear image restoration problems [3]. It utilizes a pre-trained diffusion model as a generative prior, refining only the null-space contents during the reverse diffusion process to achieve results that satisfy both data consistency and realness. This model excels in generating diverse outcomes without necessitating extra training, network modifications, or being confined to specific tasks, making it versatile across various image restoration domains. The model is described mathematically as follows:

The fundamental equation representing the degradation process is given by:

$$y = Ax \tag{4}$$

where y is the degraded image, A is the linear degradation operator, and x is the original image.

DDNM employs a unique approach by decomposing the original image x into two components as follows:

$$x' = A^{\dagger}y + (I - A^{\dagger}A)\bar{x} \tag{5}$$

Here, $A^{\dagger}$ represents the pseudo-inverse of A, and $\bar{x}$ is an arbitrary vector. This decomposition ensures that the consistency condition $Ax' \approx y$ is met, irrespective of the specific form of $\bar{x}$.

The pseudo-inverse $A^{\dagger}$ is critical for calculating the projection of y onto the range-space of A, which is expressed as $A^{\dagger}y = A^{\dagger}Ax$; The null-space component is generated by projecting $\bar{x}$ onto the null-space of A, mathematically described as $(I - A^{\dagger}A)\bar{x}$. The objective of DDNM is to find an appropriate $\bar{x}$ such that the restored image $\hat{x}$, synthesized by combining the range-space and null-space components, conforms to the distribution of natural images $q(x)$. This is symbolized as:

$$\hat{x} \sim q(x) \tag{6}$$

Through this meticulous synthesis process, employing diffusion models to generate the null-space component, DDNM ensures that the restored image not only adheres to the original degraded observation y but also embodies the natural realism expected of high-quality images. This innovative approach exemplifies how DDNM seamlessly blends consistency with realness, setting a new benchmark in the domain of image restoration.

The choice of the linear degradation operator A is crucial and varies with the image restoration task. In the task of colorization, the linear degradation operator A is chosen to convert a color image into grayscale by averaging the RGB channels. Using R, G, B to represent the red, green, and blue channels of the color image x, respectively, and y as the resulting grayscale image, the process is described mathematically as:

$$y = Ax = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{7}$$

For deblurring, the degradation operator A encapsulates the effect of a Gaussian blur. The mathematical representation of the deblurring process is not a simple matrix multiplication but implemented as a convolution with a Gaussian kernel.

For inpainting, assuming M is the binary mask with ones indicating known pixels and zeros for the pixels to be inpainted, and $\odot$ denotes the element-wise multiplication. the operator A is defined by a mask matrix that indicates the known and unknown pixel locations:

$$y = Ax = M \odot x \tag{8}$$

3.2 Stable Diffusion and Textual Prompt

Stable Diffusion is a state-of-the-art text-to-image synthesis model that employs a combination of a Variational Autoencoder (VAE) and the CLIP model to generate detailed images from textual descriptions. Initially, it uses a VAE to encode images into a latent space, where the entire

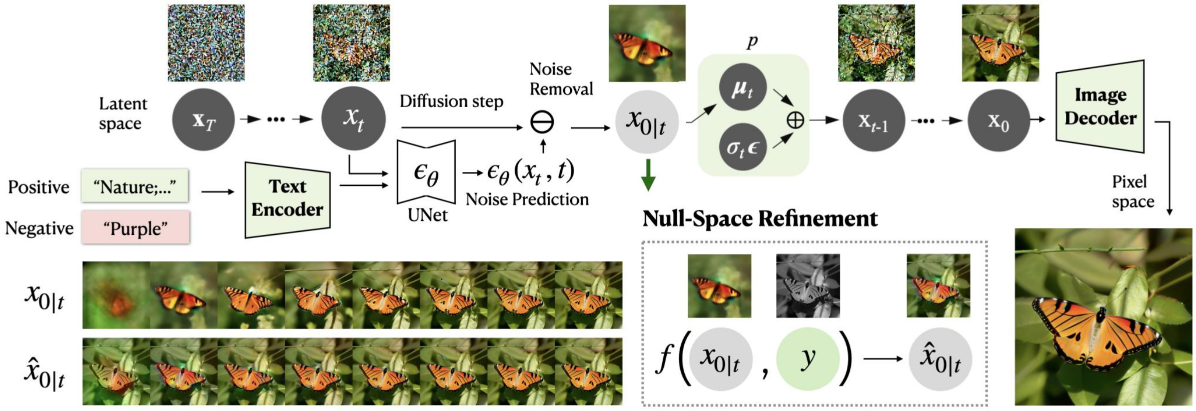

Fig. 2. Proposed pipeline integrating DDNM and Stable Diffusion.

denoising diffusion process is conducted, efficiently translating complex text embeddings into visual content. These text embeddings are produced by the CLIP model, which captures the semantic nuances of the input text, ensuring the generated images closely align with the provided descriptions. Throughout this process, a UNet-based architecture iteratively refines the latent representation, gradually removing noise to reveal the final image. Finally, an image decoder, part of the VAE, is used to convert the processed latent representation back into a high-fidelity image. This innovative approach allows Stable Diffusion to effectively bridge the gap between textual prompts and visual imagery, opening up new avenues for creative and automated content generation in various fields.

Textual prompts are integral in guiding the image restoration tasks within Stable Diffusion models. We utilize positive prompts to direct the model towards desired features and negative prompts to avert unwanted characteristics. This dichotomy is orchestrated via cross-attention mechanisms, which selectively emphasize pertinent aspects of the latent space as influenced by the prompts. Mathematically, if L_+ is the latent space swayed by positive prompts and L_- by negative prompts, the resultant latent vector L_r for image restoration is:

$$L_r = L_+ - L_- \tag{9}$$

This subtraction captures the desired characteristics to retain and those to eliminate, directing the denoising diffusion process to refine the restoration in accordance with the prompts, enhancing both visual and semantic alignment with the input text.

3.3 Proposed Pipeline

Our proposed pipeline is illustrated in Figure 2, in which we integrate the core aspects of Stable Diffusion with the Denoising Diffusion Probabilistic Model (DDPM) framework, while incorporating the Denoising Diffusion Null-Space Model (DDNM) specifically during the reverse diffusion process of $x_{0|t}$. Notably, since Stable Diffusion operates predominantly within the latent space, each denoising step entails a conversion process where $x_{0|t}$ is initially transformed from the latent to the image space for processing. Subsequently, the outcome of the DDNM operation is converted back into the latent space. This method ensures that each iteration refines the image representation while maintaining the integrity and coherence of the latent space transformations, thereby enhancing the final image quality and relevance to the input text prompts.

The following Algorithm 1 shows the whole reverse diffusion process of out method based on DDNM. Here the ϵ_θ is the UNet model with parameters θ for noise prediction in each diffusion ste

Algorithm 1 Sampling of DDNM

$x_T \sim \mathcal{N}(0, I)$
for $t = T, \ldots, 1$ **do**
  $x_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \epsilon_\theta(x_t, t)\sqrt{1 - \bar{\alpha}_t})$
  $\hat{x}_{0|t} = A^\dagger y + (I - A^\dagger A)x_{0|t}$
  $x_{t-1} \sim p(x_{t-1} \mid x_t, \hat{x}_{0|t})$
end for
return x_0

4 Analysis & Evaluation & comparison to other methods

4.1 Performance Metrics

In the evaluation of image restoration quality, two fundamental metrics are employed: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). PSNR is a widely recognized metric that estimates image reconstruction quality based on the mean squared error between the restored and original images. It is defined in terms of the logarithmic scale of the ratio between the maximum possible pixel value and the mean squared error, providing a measure of the peak error. High PSNR values indicate a higher fidelity of restoration. SSIM, on the other hand, is used to assess the perceived quality of the restored image by measuring the similarity in structural information, luminance, and contrast between the original and restored images. SSIM values range from -1 to 1, with higher values

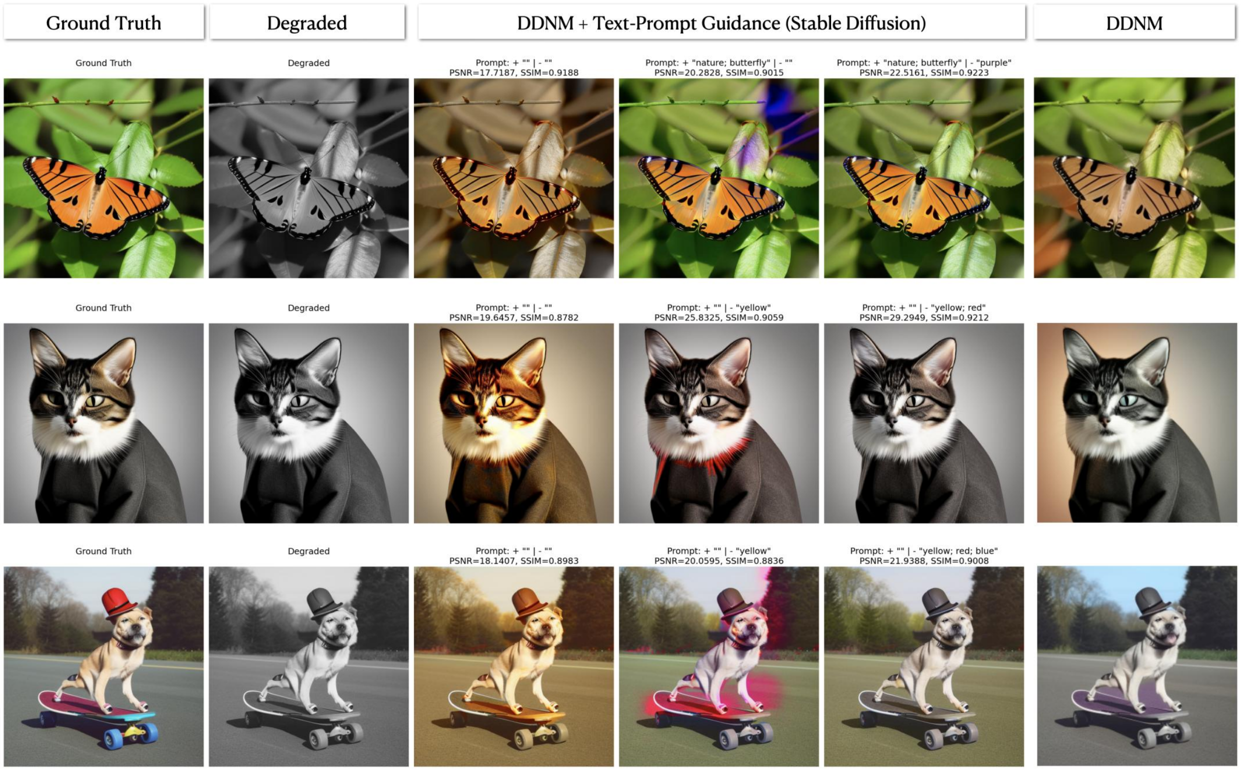

Fig. 3. The comparative colorization results of our approach using the Stable Diffusion pipeline with tailored prompts against the original DDNM framework. From left to right: the first column presents the ground truth images; the second column shows the degraded grayscale versions; the next three columns depict our method's colorized outputs, influenced by various positive and negative text prompts, which are specified above each image. The corresponding PSNR and SSIM metrics for our generated images demonstrate that with appropriately selected prompts, there is a notable enhancement in both fidelity and structural similarity to the ground truth. The final column showcases the colorized results achieved using the original DDNM method, allowing for a visual and quantitative comparison. Our method's ability to fine-tune colorization through text prompts leads to improved PSNR and SSIM values, suggesting a significant advancement over traditional DDNM techniques.

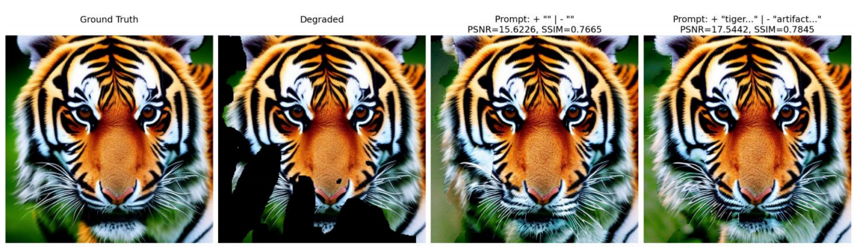

Fig. 4. The inpainting results of our approach using the Stable Diffusion pipeline with tailored prompts. For the better results with prompts, we utilize 'tiger ear; tiger face;gradient green backgrounde; left down corner tiger body' for positive prompts and 'artifact; dark green artifact;unrealistic background' for negative prompts.

signifying greater structural similarity to the ground truth image. Both PSNR and SSIM are essential in our analysis as they collectively provide insights into the accuracy of pixel values as well as the structural and textural consistency of the restored images, which are crucial aspects of visual perception and overall image quality.

4.2 Quantitative Analysis

Our approach yields quantitative enhancements in image restoration tasks, as evidenced by elevated PSNR and SSIM values upon the meticulous fine-tuning of textual prompts. By employing negative prompts, we efficiently mitigate unwanted color artifacts, whereas positive prompts are utilized to reinforce the fidelity of the content. Notably, our methodology also facilitates the generation of higher-resolution images (512x512), surpassing the original DDNM's capability of producing images at 256x256. This increase in resolution enables a more detailed and nuanced restoration, which is critical for practical applications.

4.3 Qualitative Analysis

Qualitatively, our model demonstrates a marked improvement in the reduction of artifacts when compared to the original DDNM framework. The ability to manually fine-tune the prompts provides significant leverage in optimizing the restored image's quality. For instance, in the task of colorizing a butterfly image, the incorporation of relevant positive prompts, such as "nature" and "butterfly," allows the model to learn the appropriate color palette, resulting in a more vibrant and authentic restoration. Additionally, the strategic use of negative prompts like "purple" successfully removes extraneous colors from the image background, further refining the restoration process. In the inpainting task featuring a tiger, the original outputs exhibited noticeable artifacts when no prompts were applied. However, with the inclusion of positive prompts relating to tigers and negative prompts aimed at artifacts, the resulting images were notably better and exhibited a smoother texture.

4.4 Limitations

Despite the promising results, our method is not without its limitations. The process necessitates manual intervention to fine-tune prompts, which requires careful observation and multiple adjustments. This can lead to variable performance and may not be as robust as desired in all scenarios. The efficacy of our zero-shot DDNM approach is heavily contingent upon the diffusion model's inherent generative quality. As such, the model's performance can be unpredictable, particularly in cases where the relationship between the textual prompts and the visual content is not straightforward or when the domain of the content is outside the model's training data.

5 Results

5.1 Experimental Setup

Our experimental framework is designed to rigorously evaluate the performance of our proposed image restoration technique. To ensure reproducibility and accessibility, we opted to conduct our experiments within the Google Colaboratory environment. The core of our development leverages the PyTorch library, integrate functionalities from the Hugging Face's Diffusers library, specifically exploiting the capabilities of the Stable Diffusion pipeline (version 1.4). This choice was motivated by the pipeline's proven efficacy in image generation tasks, providing a solid foundation for our image restoration endeavors.

Our model implementation draws inspiration from the DDNM framework, adapting its principles to the context of latent diffusion models. Unlike the traditional application of DDNM, we developed our pipeline independently, allowing for a seamless fusion with the latent diffusion model provided by the Stable Diffusion pipeline. This integration is critical for harnessing the full potential of the diffusion process in a zero-shot learning setup, catering to the unique demands of image restoration tasks.

The implementation of our method, along with a demonstrative Colab notebook, is openly accessible at the following URL: Google Colab.

5.2 Dataset

For the evaluation of our model, we focus on a test dataset comprising images of size 512x512 pixels. In line with our objective to streamline the experimental process and maintain control over the degradation variables, we generate the test images internally. This approach not only simplifies the experimental setup but also ensures that the degradation processes applied to the images are consistent and accurately reflect the scenarios for which our model is designed. Since we have limited time and computational resources, our experiments including colorization, deblurring and inpainting are featured on a few representative images, as shown in Figure 1, 3 and 4.

6 Discussion, limitations, future work, and conclusion

In summary, our project contributes to the evolving field of text-to-image synthesis by exploring the integration of text prompts with Stable Diffusion and DDNM for image restoration. We have taken initial steps toward correcting errors in generated images using textual descriptions, showing potential to enhance the fidelity and relevance of the restored images. While our results are promising, we recognize the limitations of our current approach and the need for further validation and improvement.

Regarding the limitation, the quality of generated images heavily relies on the training data of the Stable Diffusion model, which may not encompass all types of imagery or textual nuances. Additionally, the conversion between latent and image spaces, despite being optimized, still introduces information loss. Another limitation is the computational demand, primarily due to the iterative nature of diffusion models, which could hinder real-time applications.

Future research directions could prioritize enhancing efficiency and lowering computational overhead to enable real-time application capabilities. Investigating different architectures or optimization strategies might also mitigate information loss during the conversion between spaces. Additionally, although our current framework exhibits potential in handling noise-free environments, it remains untested in noisy conditions. Evaluating our model under various noisy scenarios could significantly underscore the robustness and applicability of our method, especially leveraging the advanced features of DDNM+. Thorough examination and validation in environments affected by diverse noise and distortions constitute essential forthcoming steps.

In conclusion, our project offers an exploration into the integration of text prompts with Stable Diffusion and DDNM for image restoration, recognizing the complexities and incremental nature of advancements in text-to-image synthesis. Our initial findings suggest a potential pathway for enhancing image accuracy with textual descriptions, yet we acknowledge the modest scope of our current investigation and the expansive room for improvement. This work is merely a stepping stone in the vast landscape of AI-driven image generation, highlighting our commitment to continuous exploration and humility in the face of technological challenges and innovations.

References

[1] J. Su, B. Xu, and H. Yin, "A survey of deep learning approaches to image restoration," *Neurocomputing*, vol. 487, pp. 46–65, 2022.

[2] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8162–8171.

[3] Y. Wang, J. Yu, and J. Zhang, "Zero-shot image restoration using denoising diffusion null-space model," *arXiv preprint arXiv:2212.00490*, 2022.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[8] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*. PMLR, 2015, pp. 2256–2265.

[9] J. Wang, Z. Yue, S. Zhou, K. C. Chan, and C. C. Loy, "Exploiting diffusion prior for real-world image super-resolution," *arXiv preprint arXiv:2305.07015*, 2023.

[10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[11] Y. Wang, J. Yu, and J. Zhang, "Zero-shot image restoration using denoising diffusion null-space model," *arXiv preprint arXiv:2212.00490*, 2022.