Prompt-Guided State-Space Foundation Model for Image Restoration

Yawei Li

The field of image restoration is continually evolving with the introduction of advanced deep learning models capable of tackling increasingly complex restoration tasks. The use of foundation models, which are pre-trained on diverse data before being fine-tuned for specific tasks, has demonstrated considerable promise in various domains of artificial intelligence. This proposal aims to develop a new foundation model for image restoration by incorporating the state-space model and enhancing it with text prompt capabilities. This approach will allow the model to perform targeted restorations based on descriptive textual prompts, significantly improving the precision and quality of the restoration process.

1 Introduction

The evolution of image restoration foundation models has been marked by significant milestones, particularly with the advent of deep learning. Initially, convolutional neural networks (CNNs) dominated the field, offering substantial improvements over traditional methods. However, the introduction of transformers revolutionized the approach to image restoration. With their ability to handle long-range dependencies and model complex patterns, transformers provided a new paradigm for addressing image degradation. Despite their success, transformers are not without limitations; they often require large amounts of data for training and can be computationally intensive, which limits their practicality for real-time applications.

In response to these challenges, the state-space model and Mamba architecture have emerged as promising alternatives. The state-space model, known for its efficiency in modeling dynamic systems, has been adapted for image restoration tasks, offering a balance between performance and computational demands. The Mamba architecture, in particular, leverages the state-space model's strengths to provide a scalable and efficient solution for image restoration, capable of handling high-resolution images and complex degradation patterns with relative ease.

The integration of textual prompts into image restoration models represents a groundbreaking shift towards more intuitive human-AI interactions. Text prompts allow users to convey their intentions and desired outcomes in natural language, making the restoration process more accessible and user-friendly. This capability is especially beneficial in scenarios where specific restoration goals are difficult to articulate through traditional interfaces, enabling a more collaborative and flexible approach to image enhancement.

The goal of this research is to develop a foundation model that harnesses the computational efficiency of the Mamba architecture and the intuitive guidance of textual prompts to set a new standard in image restoration. By combining these technologies, we aim to create a model that not only excels in restoring images but also aligns closely with user intentions, ultimately bridging the gap between advanced image processing techniques and everyday usability.

1.1 Research objectives

The research objectives for this project are as follows:

1. **To Design an Integrated System:** Develop a foundation model that seamlessly integrates the computational efficiency of the Mamba architecture with the intuitive guidance of textual prompts for image restoration tasks.
2. **To Enhance User Interaction:** Create a user-friendly interface that allows non-expert users to guide the image restoration process through natural language prompts, making the technology more accessible.
3. **To Improve Restoration Quality:** Achieve superior image restoration quality by leveraging the Mamba architecture's ability to capture long-range dependencies and complex patterns within images.

4. **To Expand Applicability:** Ensure that the model is versatile enough to handle a wide range of image restoration tasks, from common issues like noise reduction and deblurring to more complex challenges such as inpainting and super-resolution.

5. **To Optimize Computational Efficiency:** Address the computational limitations of existing models by utilizing the linear complexity of the Mamba architecture, enabling the processing of high-resolution images in a timely manner.

6. **To Validate Model Effectiveness:** Conduct extensive testing and validation to demonstrate the model's effectiveness and superiority over current state-of-the-art methods in various real-world scenarios.

7. **To Foster Collaborative Development:** Encourage collaboration within the research community by sharing insights, methodologies, and potentially the model itself, to spur further innovation in the field of image restoration.

These objectives aim to push the boundaries of what's possible in image restoration, making it more efficient, user-centric, and widely applicable.

2 Literature Review

2.1 Foundation Models in AI

Foundation models represent a paradigm shift in artificial intelligence, characterized by their ability to be adapted across a wide range of tasks and domains. These models are trained on extensive datasets and have been pivotal in powering generative AI applications such as ChatGPT, DALL-E [15, 14], and various large language models (LLMs) [19]. The adaptability of foundation models is particularly noteworthy, as they can perform tasks like natural language processing, image classification, and more with high accuracy. Despite the resource-intensive nature of their development, foundation models offer a cost-effective solution for creating specialized applications.

2.2 Image Restoration Models

Image restoration models aim to recover high-quality images from corrupted inputs, addressing issues such as noise, blur, and compression artifacts [10, 4, 16]. Recent advancements have seen the integration of diffusion models, which have shown superior performance over traditional generative adversarial network (GAN)-based methods [9, 21, 8]. These models are evaluated using benchmarks and datasets that reflect various subtasks in image restoration, such as demosaicking, spectral reconstruction, and underwater image restoration. The field continues to evolve with the development of models that can handle diverse forms of degradation.

2.3 State-Space Models and Mamba

State-space models (SSMs) have recently gained attention for their competitive performance against transformers in large-scale language modeling, while maintaining linear time and memory complexity. The Mamba architecture, a type of SSM, stands out for its linear-time sequence modeling and ability to selectively propagate information depending on the input [7]. Mamba's integration into various domains, including audio and genomics, demonstrates its versatility and potential to challenge the Transformer architecture [3].

2.4 Image Restoration with Text Prompts

The concept of textual prompt-guided image restoration is an innovative approach that introduces semantic prompts into the low-level visual domain [20]. This method allows for a natural, precise, and controllable way to perform image restoration tasks, with task-specific BERT models fine-tuned to generate textual prompts that guide the restoration process. The integration of text prompts into image restoration models like PromptIR and PromptSR has shown promising results, indicating the potential for more user-friendly and effective restoration techniques [13, 5].

3 Methodology

3.1 Problem Definition

We begin by defining the specific image restoration challenges that the model aims to address, such as noise reduction, deblurring, artifact removal, and image super-resolution. The aim of this research is to design a general model to deal with multiple image degradations with the aid of text prompts. We also need to determine the types of textual prompts that will be used to guide the restoration process and how they will be interpreted by the model.

3.2 Model Architecture Design

We will adapt the Mamba architecture to suit the needs of image restoration, incorporating any necessary modifications for efficient processing and integration with textual prompts. In addition, we also need to design a textual prompt integration mechanism that can process and utilize textual prompts to influence the restoration outcomes. Previous methods use diffusion models and cross-attention between text embeddings and image embeddings, which can be a starting point for this research.

3.3 Dataset Collection and Preparation

We will compile a comprehensive dataset of images that exhibit various forms of degradation. There are already paired dataset with ground-truth and degraded images including GoPro [12], HIDE [18]), RealBlur-R [17], and DPDD [1]. In addition to that, degraded images can also be generated from the ground-truth images including DIV2K [2], and LSDIR [11].

Besides the image dataset, we also need to create a corresponding dataset of textual prompts that describe the desired restoration actions for each image. This can be generated by large language models like GPT [6].

3.4 Model Training

We will first use the image dataset to pre-train the model on general image restoration tasks without textual prompts. Then we will introduce the textual prompts during the fine-tuning phase to teach the model how to interpret and apply the prompts to guide the restoration process.

3.5 Evaluation and Testing

Both quantitative and qualitative evaluation will be conducted:

1. **Quantitative Evaluation:** Assess the model's performance using standard image restoration metrics such as Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS).

2. **Qualitative Evaluation:** Conduct user studies to evaluate the effectiveness of the textual prompts in guiding the restoration process and the overall user experience.

4 Expected Outcomes

This project expects to develop a robust foundation model capable of contextual and high-fidelity image restoration. The model should excel in understanding and executing complex restoration tasks guided by natural language prompts, thereby providing a customizable tool for various applications.

5 Conclusion

In conclusion, this research proposal outlines a comprehensive plan to develop a foundation model for image restoration that leverages the Mamba architecture and textual prompts. The integration of these two innovative approaches promises to enhance the quality and accessibility of image

restoration techniques. The Mamba architecture's computational efficiency and the intuitive guidance provided by textual prompts are poised to set a new standard in the field. The expected outcomes include a robust foundation model, improved restoration quality, a user-friendly interface, and contributions to academic research. By achieving these goals, the project aims to revolutionize image restoration, making it more efficient and accessible to a wider range of users. The success of this research could have far-reaching implications, not only for the field of image restoration but also for the broader domain of AI, where the fusion of state-of-the-art architectures with human-like interaction is becoming increasingly important.

References

[1] A. Abuolaim and M. S. Brown. Defocus deblurring using dual-pixel data. In *European Conference on Computer Vision*, pages 111–126. Springer, 2020.

[2] E. Agustsson and R. Timofte. NTIRE 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 126–135, 2017.

[3] Q. Anthony, Y. Tokpanov, P. Glorioso, and B. Millidge. Blackmamba: Mixture of experts for state-space models. *arXiv preprint arXiv:2402.01771*, 2024.

[4] X. Chen, X. Wang, J. Zhou, Y. Qiao, and C. Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22367–22377, 2023.

[5] Z. Chen, Y. Zhang, J. Gu, X. Yuan, L. Kong, G. Chen, and X. Yang. Image super-resolution with text prompt diffusion. *arXiv preprint arXiv:2311.14282*, 2023.

[6] M. V. Conde, G. Geigle, and R. Timofte. High-quality image restoration following human instructions. *arXiv preprint arXiv:2401.16468*, 2024.

[7] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[8] B. Kawar, M. Elad, S. Ermon, and J. Song. Denoising diffusion restoration models. *Advances in Neural Information Processing Systems*, 35:23593–23606, 2022.

[9] X. Li, Y. Ren, X. Jin, C. Lan, X. Wang, W. Zeng, X. Wang, and Z. Chen. Diffusion models for image restoration and enhancement–a comprehensive survey. *arXiv preprint arXiv:2308.09388*, 2023.

[10] Y. Li, Y. Fan, X. Xiang, D. Demandolx, R. Ranjan, R. Timofte, and L. Van Gool. Efficient and explicit modelling of image hierarchies for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18278–18289, 2023.

[11] Y. Li, K. Zhang, J. Liang, J. Cao, C. Liu, R. Gong, Y. Zhang, H. Tang, Y. Liu, D. Demandolx, et al. LSDIR: A large scale dataset for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1775–1787, 2023.

[12] S. Nah, T. Hyun Kim, and K. Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3883–3891, 2017.

[13] V. Potlapalli, S. W. Zamir, S. Khan, and F. S. Khan. Promptir: Prompting for all-in-one blind image restoration. *arXiv preprint arXiv:2306.13090*, 2023.

[14] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[15] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[16] B. Ren, Y. Li, J. Liang, R. Ranjan, M. Liu, R. Cucchiara, L. Van Gool, and N. Sebe. Key-graph transformer for image restoration. *arXiv preprint arXiv:2402.02634*, 2024.

[17] J. Rim, H. Lee, J. Won, and S. Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *European Conference on Computer Vision*, pages 184–201. Springer, 2020.

[18] Z. Shen, W. Wang, X. Lu, J. Shen, H. Ling, T. Xu, and L. Shao. Human-aware motion deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5572–5581, 2019.

[19] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[20] Q. Yan, A. Jiang, K. Chen, L. Peng, Q. Yi, and C. Zhang. Textual prompt guided image restoration. *arXiv preprint arXiv:2312.06162*, 2023.

[21] Y. Zhu, K. Zhang, J. Liang, J. Cao, B. Wen, R. Timofte, and L. Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1219–1229, 2023.