High-resolution Facial Image Synthesis from Low-resolution Images and Forensic Sketches

Priyanka Kamila

Master of Science in Computer Vision, Robotics, Machine Learning

from the

University of Surrey

School of Computer Science and Electrical and Electronic Engineering

Faculty of Engineering and Physical Sciences

University of Surrey

Guildford, Surrey, GU2 7XH, UK

September 2024

Supervised by: Prof. Yi-Zhe Song

DECLARATION OF ORIGINALITY

I confirm that the project dissertation I am submitting is entirely my own work and that any material used from other sources has been clearly identified and properly acknowledged and referenced. In submitting this final version of my report to the JISC anti-plagiarism software resource, I confirm that my work does not contravene the university regulations on plagiarism as described in the Student Handbook. In so doing I also acknowledge that I may be held to account for any particular instances of uncited work detected by the JISC anti-plagiarism software, or as may be found by the project examiner or project organiser. I also understand that if an allegation of plagiarism is upheld via an Academic Misconduct Hearing, then I may forfeit any credit for this module or a more severe penalty may be agreed.

High-resolution Facial Image Synthesis from Low-resolution Images and Forensic Sketches

Priyanka Kamila

Author Signature Date:3rd September, 2024

Supervisor's name: Prof. Yi-Zhe Song, Subhadeep Koley

WORD COUNT

Number of Pages: 52

Number of Words: 14181

ABSTRACT

In this dissertation, we propose a novel dual-encoder architecture to develop person-centric image enhancement in challenging environments with low light. We utilize a dual-encoder system to process two input modalities namely sketches and low-resolution images to generate high-resolution images. These inputs are very crucial in generating distinct style vectors, which are seamlessly integrated into a pre-trained StyleGAN generator, significantly expanding the latent space for versatile image translation tasks. Unlike traditional methodologies, our approach of implementing a dual-encoder system eliminates the need for additional optimization during the embedding of real images into the extended W+ latent space. This inherently supports multi-modal synthesis through style resampling and highlighting its robustness in challenging conditions.

CONTENTS

LIST OF FIGURES

1 INTRODUCTION

As our world becomes increasingly visual, with billions of images captured and shared daily, the ability to improve image quality becomes not just a technical achievement, but a gateway to unlocking valuable information across various domains. This dissertation presents a novel approach to image enhancement that stands at the intersection of traditional computer vision techniques and cutting-edge generative models, with a particular focus on facial image processing.

Image enhancement, particularly in the context of facial images, is a complex task that has seen significant advancements in recent years. However, existing methods often struggle when confronted with degradations such as low resolution. These challenging scenarios are not merely academic exercises but reflect real-world situations encountered in surveillance, historical image restoration, and various other applications where high-quality images are crucial but often unavailable.

Our research proposes a groundbreaking solution to these challenges by integrating sketch information into the image enhancement process. This multi-modal approach leverages the complementary nature of degraded images and corresponding sketches, potentially overcoming the limitations of single-input methods. By incorporating sketches, we aim to provide additional structural and semantic information that may be lost or unclear in the low-quality input image.

By demonstrating the effectiveness of combining image and sketch inputs, this work could pave the way for more sophisticated multi-modal systems in various computer vision tasks. Our approach has the potential to improve the quality of enhanced images in scenarios where current methods fall short, expanding the practical applications of image restoration technology. The integration of sketches represents a novel way to incorporate human knowledge and perception into automated image processing systems, potentially leading to more intuitive and powerful tools for image manipulation. By addressing the ethical implications of advanced image enhancement techniques, this research contributes to the ongoing dialogue about responsible AI development in the era of deepfakes and privacy concerns. The potential applications of this technology span multiple fields, from forensics to medical imaging, highlighting the far-reaching impact of advancements in computer vision.

The integration of sketches into the image enhancement process represents a novel approach to addressing the limitations of existing methods. By leveraging the complementary information provided by sketches, we aim to overcome the challenges posed by low-resolution face inputs. This multi-modal approach has the potential to significantly improve the quality and accuracy of generated high-resolution images.

Our proposed extension to the pSp framework builds upon the foundation laid by Elad Richardson et al. (2021). The original pSp framework's innovative use of an encoder network to generate style vectors, which are then integrated into a pretrained StyleGAN generator, provides a robust starting point for our research. The resulting extended W+ latent space has demonstrated remarkable versatility in handling a wide range of image translation tasks.

To adapt this framework for our specific goals, we propose implementing a dual-encoder architecture. This architecture will consist of two parallel encoders: one dedicated to processing the low-resolution input image, and another designed to handle the corresponding sketch input. [18] By processing these inputs simultaneously, we can extract and combine features from both modalities, potentially leading to more accurate and detailed output images.

The image encoder will focus on extracting low-level features such as texture, color, and overall structure from the input image, even when it is of low quality or partially occluded. Concurrently, the sketch encoder will process the corresponding sketch, extracting high-level structural information that may be missing or unclear in the original image. [4] This dual-stream approach allows us to leverage the strengths of both input types, compensating for the weaknesses of each.

The outputs from these two encoders will be combined through a fusion module, which will learn to weigh and integrate the features from both streams optimally. This fusion process is crucial, as it determines how the information from the sketch complements and enhances the information from the image. We hypothesize that this combined representation will provide a more robust and informative basis for the subsequent image generation process.

Following the fusion of features, the combined representation will be used to generate style vectors, similar to the original pSp framework. These style vectors will then be fed into the pretrained StyleGAN generator, which will produce the final high-resolution output image. By leveraging the power of StyleGAN's ability to generate highly realistic images, we expect to achieve output images that not only have improved resolution but also maintain a high degree of realism and detail.

The successful implementation of our sketch-assisted image enhancement framework could have far-reaching implications across various domains. Law enforcement agencies could use this technology to enhance low-quality surveillance footage, potentially aiding in suspect identification.

In medical contexts, the system could be used to enhance low-resolution or noisy medical images, with expert-drawn sketches highlighting areas of interest. [9] The fashion and beauty industries could leverage this technology for virtual try-on applications, using sketches to guide the placement of virtual makeup or accessories on user-uploaded photos. The system could be developed into an assistive technology for visually impaired individuals, converting low-quality or abstract visual information into more clear, high-resolution images.

Future research directions could explore the extension of this framework to other domains beyond face images, such as full-body poses or even non-human subjects. Additionally, investigating the potential of interactive sketch refinement, where the system provides real-time feedback as users adjust their sketches, could lead to more intuitive and powerful image enhancement tools.

As with any technology that involves the manipulation and enhancement of images, particularly those of human faces, it is crucial to consider the ethical implications of our research. The ability to enhance low-quality images may raise privacy concerns, especially in the context of surveillance footage. Clear guidelines and regulations must be established to prevent misuse. The technology could potentially be misused to create or enhance fake images for the purpose of spreading misinformation. Developing robust detection methods for images generated by our system will be crucial.

By addressing these ethical considerations head-on and implementing appropriate safeguards, we can ensure that our research contributes positively to the field of computer vision while minimizing potential negative impacts on society.

In conclusion, our proposed extension to the pSp framework, incorporating sketch assistance for image enhancement, has the potential to significantly advance the state of the art in image restoration and manipulation. By leveraging multi-modal inputs and the power of modern generative models, we aim to push the boundaries of what is possible in challenging image enhancement scenarios. As we progress with this research, we remain committed to rigorous evaluation, ethical consideration, and exploration of diverse applications that can benefit from this technology.

1.1 Background and Context

In the realm of computer vision and image processing, the ability to handle diverse input modalities and perform complex image-to-image translation tasks has become increasingly important. This research proposes an innovative approach that leverages a dual-encoder architecture to address challenges simultaneously. By developing a unified framework capable of processing these varied inputs, we aim to push the boundaries of image enhancement and translation techniques, with a particular focus on facial image processing.

1.1.1 Ethical Considerations

As with any technology that involves the manipulation and enhancement of images, particularly those of human faces, it is crucial to consider the ethical implications of our research. The ability to enhance low-quality images may raise privacy concerns, especially in the context of surveillance footage. The training dataset used for building the model is diverse and representative, to avoid perpetuating or amplifying biases in the generated images. The technology could potentially be misused to create or enhance fake images to spread misinformation. [2] Developing robust detection methods for images generated by our system will be crucial.

In conclusion, our proposed extension to the pSp framework, incorporating sketch assistance for image enhancement, has the potential to significantly advance the state of the art in image restoration and manipulation. By leveraging multi-modal inputs and the power of modern generative models, we aim to push the boundaries of what is possible in challenging image enhancement scenarios. As we progress with this research, we remain committed to rigorous evaluation, ethical consideration, and exploration of diverse applications that can benefit from this technology.

1.2 Objectives

The core of this research revolves around developing and integrating a dual-encoder architecture. This novel approach allows for the simultaneous processing of multiple input types, specifically low-light images and sketch representations. By using two encoders working in tandem, the system can effectively capture and process the unique characteristics of each input modality, providing a more robust and versatile foundation for image-to-image translation tasks.

To accommodate the extended latent space resulting from the dual-encoder architecture, the training process must be carefully designed and optimized. [3] This involves developing streamlined optimization techniques that can effectively handle the increased complexity of the model while maintaining efficiency. The focus is on creating a training regimen that allows the system to learn and generalize across various input types and translation tasks.

The proposed framework will be rigorously tested on facial image-to-image translation tasks. [5] This validation process aims to demonstrate the system's superior performance compared to existing task-specific solutions. By showcasing its ability to handle multiple input modalities and produce high-quality output across various facial translation tasks, we can establish the effectiveness and versatility of the dual-encoder approach.

While the initial focus is on facial image processing, the research will investigate the adaptability of the framework to domains beyond human faces. This exploration will assess the potential for broader image enhancement scenarios, testing the system's capability to generalize across different types of images and translation tasks. This step is crucial in determining the wider applicability and impact of the proposed architecture. [16]

The efficiency of the dual-encoder system will be thoroughly assessed across a range of image translation challenges. This evaluation will focus on demonstrating the framework's superiority in handling diverse tasks, comparing its performance, speed, and resource utilization against existing single-purpose solutions. The goal is to showcase how a unified approach can outperform specialized models across multiple domains.

A key feature of the proposed framework is its inherent support for multi-modal synthesis. Through style resampling techniques, [11] the system can accommodate various input modalities and generate diverse outputs. This versatility will be demonstrated, highlighting the architecture's ability to adapt to different input types and produce a range of stylized outputs, further showcasing its flexibility and potential applications.

- To develop and integrate a dual-encoder architecture to simultaneously process low resolution and sketch input images
- Enhancing the training process to accommodate the extended latent space, focusing on streamlined optimization

- Validating the idea with testing on facial image-to-image translation tasks
- Investigating the adaptability of the framework beyond human facial domains, exploring it's potential for broader image enhancement scenarios
- Assessing the efficiency of the dual-encoder system in handling tasks demonstrating its superiority in diverse image translation challenges
- Demonstrating the inherent support for multi-modal synthesis through style resampling, showcasing its versatility in accommodating various input modalities

The proposed dual-encoder architecture represents a significant advancement in the field of image-to-image translation and enhancement. By addressing multiple challenges simultaneously and demonstrating superior performance across various tasks, this research has the potential to revolutionize how we approach complex image processing problems. The framework's ability to handle diverse input modalities coupled with its efficiency and adaptability opens up new possibilities for applications in facial image processing and beyond. As we continue to validate and refine this approach, we anticipate that it will contribute significantly to the development of more versatile and powerful image manipulation tools, benefiting areas such as computer vision, augmented reality, and content creation.

2 BACKGROUND THEORY AND LITERATURE REVIEW

There has been significant progress in computer vision, notably in areas of image synthesis, manipulation, and analysis, owing to the advancements in deep learning techniques. The invention of Generative Adversarial Networks (GANs) is most often marked as a milestone as it has enabled systems to generate, edit, and interpret visual data, opening up new avenues and giving creative control over generating style content in various industries.

Our work is particularly focused on applications in the field of forensics, where the ability to generate accurate and detailed visual representations from diverse and often incomplete inputs can be of crucial importance. By extracting and combining descriptors from the two distinct input modalities: low-resolution photographic images and hand-drawn sketches, our system will aim to harness the strengths of both input types. Low-resolution facial input images will provide real-world visual information, while sketches will offer the flexibility to capture abstract details that might not be present in available photographs.

We will proceed to intelligently mix and match the descriptors extracted from these inputs and pass them onto a pre-trained StyleGAN decoder. This approach leverages the power of StyleGAN to generate high-resolution images. The result is a flexible and powerful system capable of generating highly detailed and accurate visual representations from a combination of photographic and sketch-based inputs. Law enforcement agencies often work with incomplete visual information when trying to identify suspects or missing individuals. Our system could aid in generating more accurate facial composites, age progression or regression visualizations, and even reconstructions based on partial photographic evidence combined with witness descriptions translated into sketches. However, the implications of this research extend beyond forensic applications. The ability to seamlessly combine different types of visual inputs in generating high-quality images could have far-reaching impacts in fields such as entertainment, design, and virtual reality.

In this thesis, we will explore the theoretical foundations underpinning our approach, drawing from a rich body of literature in deep learning, computer vision, and forensic science. We will then detail the architecture and implementation of our dual encoder system, followed by a comprehensive evaluation of its performance across various tasks and scenarios. As we dive deeper into this exciting intersection of technologies, we aim not only to present a novel technical approach but also to critically examine its potential impacts, both positive and negative. We will discuss the ethical considerations surrounding such powerful image-generation capabilities and explore potential safeguards against misuse. Through this work, we hope to contribute to the ongoing advancement of AI-assisted image synthesis and analysis, with a particular focus on applications that can aid in critical real-world scenarios such as criminal investigations and missing person cases. By

bridging the gap between different visual input modalities and leveraging state-of-the-art generative models, we aim to push the boundaries of what's possible in AI-assisted visual reconstruction and synthesis.

2.0.1 Foundational Concept of GANs

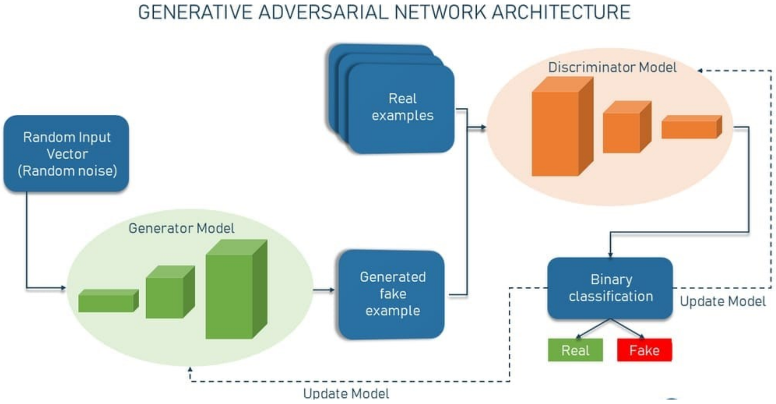


Figure 2.1: GAN Network architecture

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. in 2014, represent a revolutionary approach to generative modeling in machine learning. The core concept of GANs involves two neural networks—a generator and a discriminator—engaged in an adversarial training process. This framework has proven remarkably effective for generating high-quality, realistic data, particularly in the domain of image synthesis.

The generator network in a GAN aims to produce synthetic data that is indistinguishable from real data. It takes random noise as input and transforms it into data samples, such as images. The discriminator network, on the other hand, acts as a binary classifier, attempting to distinguish between real data samples and those produced by the generator.

During training, these two networks are pitted against each other in a minimax game. The generator tries to produce increasingly realistic data to fool the discriminator, while the discriminator works to improve its ability to detect synthetic samples. This adversarial process continues until an equilibrium is reached, where the generator produces data so realistic that the discriminator can no longer reliably distinguish it from real data.

The power of GANs lies in their ability to learn complex data distributions without explicit probability density estimation. This makes them particularly well-suited for tasks involving high-dimensional data, such as images. GANs have demonstrated remarkable success in generating photorealistic images, performing style transfer, and even in domain adaptation tasks.

However, training GANs can be challenging due to issues such as mode collapse (where the generator produces a limited variety of samples) and training instability. Numerous variations and improvements have been proposed to address these challenges, leading to more stable and controllable GAN architectures.

Building upon the foundational GAN framework, StyleGAN, introduced by Karras et al. in 2019, represents a significant leap forward in the quality and controllability of generated images. StyleGAN introduces several key innovations that have made it a cornerstone in state-of-the-art image synthesis.

The primary innovation of StyleGAN is its style-based generator architecture. Unlike traditional GANs where the input latent code is fed directly into the generator, StyleGAN introduces a mapping network that transforms the input latent code into an intermediate latent space ’w’. This intermediate latent space is then used to control the ’styles’ at each convolution layer of the generator through adaptive instance normalization (AdaIN).

This architecture allows StyleGAN to separate high-level attributes (such as pose, identity, or gender in face generation) from stochastic variation (like hair, freckles, or background details). This separation provides unprecedented control over the generated images and allows for smooth interpolation between different styles.

Another key feature of StyleGAN is its progressive growing approach. The network starts by generating low-resolution images and progressively increases the resolution by adding new layers. This technique allows for more stable training and enables the generation of high-resolution images with remarkable detail.

StyleGAN also introduces a mixing regularization technique, where two latent codes are mixed to create an image, encouraging the network to localize styles to appropriate levels of the generator hierarchy. This further enhances the network's ability to disentangle different aspects of the generated images.

The success of StyleGAN led to further improvements in StyleGAN2, which addressed issues like water droplet-like artifacts in the original model and introduced path length regularization for more stable and higher-quality results.

2.0.2 Image-to-Image Translation and the pSp Framework

As GANs evolved, researchers began exploring their potential for more controlled image generation tasks, particularly in the domain of image-to-image translation. This field focuses on learning the mapping between input and output images, often from different domains or with different characteristics.

Early works in this area, such as Pix2Pix and CycleGAN, demonstrated the potential of GANs for tasks like colorization, style transfer, and domain adaptation. However, these methods often struggled with preserving fine details and maintaining consistency across different tasks.

The introduction of the Pixel2Style2Pixel (pSp) framework by Richardson et al. [13] in 2020 marked a significant advancement in this field. pSp bridges the gap between real images and StyleGAN's latent space, enabling high-fidelity image-to-image translation tasks.

The key innovation of pSp is its ability to directly encode real images into StyleGAN's extended latent space w+. This is achieved through a novel encoder architecture that combines a feature pyramid with map2style blocks. The feature pyramid allows the encoder to capture information at multiple scales, while the map2style blocks learn to map these features directly to StyleGAN's w+ space.

By leveraging the rich representational power of StyleGAN's latent space, pSp enables a wide range of image manipulation tasks, including inversion (reconstructing an input image), face frontalization, conditional image synthesis, and more. The framework demonstrates superior performance compared to previous methods, particularly in preserving fine details and achieving photorealistic results.

The pSp framework's ability to work with StyleGAN's latent space also allows for easy integration of other StyleGAN-based manipulation techniques, making it a versatile tool for various image editing and synthesis tasks.

Building upon the success of single-encoder models like pSp, researchers have explored the potential of dual encoder architectures to handle multiple input modalities or to capture different aspects of the input data. This approach is particularly relevant for tasks that involve combining information from diverse sources, such as photographs and sketches.

Dual encoder architectures typically consist of two separate encoding pathways, each specialized for a particular type of input. These encoders can be trained to project their respective inputs into a shared latent space, allowing for the combination and manipulation of features from both sources.

In the context of image synthesis, dual encoders have been applied to tasks such as cross-modal retrieval, where one encoder processes text descriptions and another processes images, allowing for text-to-image or image-to-text matching. Similarly, in the domain of face synthesis, dual encoder approaches have been used to separately capture identity and attribute information, allowing for more controlled generation and editing of facial images.

Sketch-based image synthesis represents a particularly challenging and relevant application of these techniques. Sketches provide a flexible and intuitive way for users to specify desired image characteristics, but translating these abstract, often incomplete representations into photorealistic images poses significant challenges.

Recent advancements in this area have leveraged deep learning techniques to bridge the gap between sketches and photorealistic images. Approaches like SketchyGAN have demonstrated the feasibility of generating diverse and realistic images from simple sketches, opening up new possibilities for user-guided image generation and editing.

The combination of dual encoder architectures with sketch-based synthesis techniques offers powerful capabilities for image generation systems. By allowing the integration of both photographic and sketch-based inputs, such systems can leverage the strengths of both modalities – the detail and realism of photographs, and the flexibility and user control offered by sketches.

2.0.3 Feature Mixing and Style Manipulation

The ability to mix features and manipulate styles is crucial for creating flexible and controllable image generation systems. This aspect of the technology builds upon the rich latent representations learned by models like StyleGAN and the encoding capabilities of frameworks like pSp.

Feature mixing involves combining or interpolating between the latent representations of different inputs. In the context of StyleGAN, this can be achieved by mixing the style codes at different levels of the generator, allowing for fine-grained control over various aspects of the generated image.

Style manipulation, on the other hand, involves identifying and modifying specific directions in the latent space that correspond to particular image attributes or styles. Techniques like InterFaceGAN and GANSpace have demonstrated that it's possible to discover interpretable directions in the latent space that correspond to semantic attributes like age, gender, or facial expression.

By combining feature mixing and style manipulation techniques, it becomes possible to create highly controllable image generation systems. These systems can blend characteristics from multiple inputs, transfer styles between images, or apply specific attribute changes while maintaining the overall identity and structure of the input.

In the context of a dual encoder system processing both photographic and sketch inputs, these techniques allow for sophisticated combinations of information from both sources. For example, one could use the overall structure and identity information from a photograph, but incorporate specific details or modifications specified in a sketch.

2.0.4 High-Resolution Image Generation

The ability to generate high-resolution images is crucial for many practical applications, particularly in fields like forensics where fine details can be of utmost importance. However, generating high-resolution images poses several challenges, including increased computational requirements and the need to maintain coherence across larger spatial scales.

Recent advancements in GAN architectures, particularly StyleGAN, have made significant strides in high-resolution image generation. The progressive growing approach used in StyleGAN, where the network gradually increases the resolution of generated images during training, has proven effective in producing detailed, high-resolution outputs.

Furthermore, techniques like super-resolution have been developed to enhance the resolution of generated or existing images. These methods often use GANs or other deep learning approaches to intelligently upsample images, inferring plausible high-frequency details.

In the context of a dual encoder system for forensic applications, the ability to generate high-resolution outputs is crucial. It allows for the production of detailed facial composites or reconstructions that can be more useful in identification tasks. The challenge lies in effectively combining the information from potentially low-resolution or incomplete inputs (like sketches or partial photographs) to produce a high-resolution, detailed output.

2.0.5 Forensic Applications

The convergence of advanced image generation techniques, dual encoder architectures, and high-resolution synthesis opens up exciting possibilities in the field of forensic science. These technologies have the potential to revolutionize how visual evidence is processed and utilized in criminal investigations.

One of the most prominent applications is in the generation of facial composites. Traditional methods of creating facial composites often rely on assembling features from a limited set of options, resulting in unrealistic or inaccurate representations. With GAN-based approaches, it becomes possible to generate highly realistic and detailed facial images based on descriptions or partial information.

The dual encoder approach, combining photographic and sketch inputs, is particularly powerful in this context. It allows for the integration of available photographic evidence with sketch-based inputs that can capture remembered details or witness descriptions. This combination can potentially lead to more accurate and useful facial reconstructions.

Age progression and regression represent another important application area. By manipulating the appropriate latent space directions, it's possible to generate plausible representations of how an individual might look at different ages. This can be crucial in cases involving missing persons or long-term fugitives.

These technologies also have potential applications in forensic reconstruction, allowing for the generation of complete facial images from partial remains or damaged photographs. The ability to work with multiple input modalities and generate high-resolution outputs makes these systems particularly well-suited for such challenging tasks.

However, the power of these technologies also raises important ethical considerations. The ability to generate highly realistic fake images could potentially be misused for deception or fraud. As such, it's crucial that the development of these technologies is accompanied by robust safeguards and ethical guidelines for their use in forensic contexts.

In conclusion, the integration of GANs, StyleGAN, dual encoder architectures, and advanced image synthesis techniques presents a powerful toolkit for forensic applications. By enabling the generation of high-quality, detailed images from diverse and often incomplete inputs, these technologies have the potential to significantly enhance the capabilities of forensic investigators. As research in this field continues to advance, we can anticipate even more sophisticated and accurate image generation systems, further revolutionizing the field of forensic visual analysis.

2.1 Literature Review

Building upon the success of StyleGAN, researchers have developed various techniques to leverage its powerful latent space for image manipulation tasks. One such technique is presented in by Abdal et al. [1]. This work tackles the challenge of embedding real images into StyleGAN's latent space, introducing an optimization-based method to find the best latent code that reconstructs a given image. By demonstrating that using an extended latent space (W+) allows for better reconstruction quality than the original W space, Image2StyleGAN enables various image manipulation tasks by leveraging StyleGAN's rich latent representation.

The ability to manipulate images in the latent space of GANs has opened up new possibilities for semantic editing and attribute manipulation. Two notable works in this area are ”Interpreting the Latent Space of GANs for Semantic Face Editing” by Shen et al. (2020) [14] and [7] by Härkönen et al. (2020).

Shen et al. proposed a framework called InterFaceGAN, which discovers that after linear transformations, the latent codes of well-trained GANs exhibit a disentangled representation of various facial attributes. This discovery allows for precise control over attributes like gender, age, expression, and pose, enabling more granular and intuitive image editing.

Härkönen et al. introduce GANSpace, a method for discovering interpretable controls within the latent space of GANs. By analyzing the latent space, GANSpace identifies directions that correspond to meaningful semantic attributes, providing greater control over the generated output and offering insights into the underlying structure of GAN latent spaces.

As GANs evolved, researchers began to explore their potential for image-to-image translation tasks. A significant breakthrough in this area came with the introduction of the Pixel2Style2Pixel (pSp) encoder by Richardson et al. [13]

The pSp encoder represents a cornerstone in the field of image-to-image translation, bridging the gap between real images and StyleGAN's latent space. By mapping real images directly into StyleGAN's W+ latent space, pSp enables high-fidelity image-to-image translation tasks that leverage StyleGAN's powerful generation capabilities.

The pSp architecture combines a feature pyramid with map2style blocks, allowing for fine-grained control over generated images. This novel approach demonstrates superior performance in various tasks such as inversion, face frontalization, and conditional image synthesis. By effectively bridging the gap between real images and StyleGAN's latent space, pSp has opened up new possibilities for image manipulation and synthesis.

The concept of dual encoders has emerged as a powerful approach for handling multiple input modalities or styles. This approach has been explored in various contexts, demonstrating the benefits of using complementary encoders to capture different aspects of input data.

One notable example is the work by Zhao et al. (2017) [19] introduces a system that employs two generative agents, one focusing on identity preservation and the other on photorealism. Through adversarial training, these agents collaboratively learn to synthesize high-quality profile faces that accurately capture an individual's identity while maintaining realistic visual details.

Another significant contribution to this area is the paper [8].This work proposes a dual auto-encoder model with bidirectional latent space regression for robust image translation and completion. The approach allows the model to separate different facial attributes, such as identity, expression, and pose, into distinct latent factors, enabling various tasks including facial image editing, generation, and attribute manipulation.

The integration of multiple input modalities, such as images and sketches, relates closely to the field of multi-modal learning. Wang et al.'s work [17] demonstrates the effectiveness of dual-branch networks for cross-modal tasks, providing insights that can be applied to systems combining diverse input types like photographs and sketches.

The incorporation of sketch inputs in image synthesis and retrieval systems represents an important development, particularly for applications in forensics and creative fields. Yu et al. (2016) [12]addresses the task of retrieving images from a database based on user-provided sketches.

This work proposes a novel method utilizing convolutional neural networks (CNNs) with a two-stream architecture - one for processing sketches and another for handling images. Both streams extract deep features that capture underlying visual information, and a joint embedding space is learned to bridge the gap between sketches and images. This approach demonstrates impressive performance in matching sketches to corresponding images, even with significant variations in style, pose, and viewpoint.

Building upon this foundation, Chen and Hays 2018, [4] further advancing the field of sketch-based image generation. These works collectively showcase the potential of deep learning techniques in translating abstract, user-generated sketches into realistic images - a capability with significant implications for forensic applications.

The advancements in GAN technology, image-to-image translation, and sketch-based image generation have opened up new possibilities in the field of forensic science. Two papers are particularly relevant in this context:

Ichim et al. explores the application of deep learning to the challenging task of creating realistic facial images from skull remains. Their approach utilizes a deep convolutional neural network (CNN) to learn complex patterns and relationships between skull features and corresponding facial images. The results demonstrate that deep learning models can significantly outperform traditional methods, producing more accurate and realistic facial reconstructions.

Complementing this, Frowd's 2021 survey paper, ”Deep Learning Approaches for Facial Composite Generation,” provides a comprehensive overview of various techniques, including GANs, variational autoencoders (VAEs), and hybrid models, for generating facial composites. This work highlights the potential of deep learning methods to revolutionize forensic facial reconstruction and composite generation, offering law enforcement agencies powerful new tools for investigation and identification.

StyTr2 revolutionizes image style transfer by harnessing the capabilities of transformer architectures. Unlike traditional CNN-based methods [6] introduces the approach of embracing long-range dependencies, crucial for maintaining global image information during style transfer. StyTr2 incorporates two distinct transformer encoders, each responsible for generating domain-specific sequences for content and style, respectively. Subsequently, a multi-layer transformer decoder stylizes the content sequence based on the style sequence. Furthermore, we address the limitations of existing positional encoding methods by introducing content-aware positional encoding (CAPE), offering scale-invariant properties ideal for image style transfer tasks. Through qualitative and quantitative experiments, we showcase the effectiveness of StyTr2 compared to state-of-the-art CNN-based and flow-based approaches.

The paper [10] proposes a novel method for image translation and completion by using a dual auto-encoder architecture with bidirectional latent space regression. This approach allows for robust and accurate image transformation between different domains, even in the presence of missing or corrupted data. The model learns a shared latent space that captures the underlying structure of both input and output domains, enabling effective mapping and reconstruction.

A dual auto-encoder model with bidirectional latent space regression is proposed for robust image translation and completion in the paper. This approach allows the model to separate different facial attributes, such as identity, expression, and pose, into distinct latent factors. The disentangled representation can be used for various tasks, including facial image editing, generation, and attribute manipulation.

Zhao et al. (2017) [15] introduce Dual-Agent GANs for photorealistic and identity-preserving profile face synthesis. The paper employs two generative agents, one focusing on identity preservation and the other on photorealism. By leveraging adversarial training, the agents collaboratively learn to synthesize high-quality profile faces that accurately capture the individual's identity while maintaining realistic visual details. This approach outperforms existing methods in terms of both identity preservation and visual quality, enabling more realistic and personalized face generation.

2.2 Sketch-Based Image translation

The authors [12]propose a novel method utilizing convolutional neural networks (CNNs) to achieve this. The model employs a two-stream architecture, where one stream processes sketches and the other handles images. Both streams extract deep features that capture the underlying visual information. To bridge the gap between sketches and images, a joint embedding space is learned, allowing for effective mapping and comparison. The model demonstrates impressive performance, successfully matching sketches to corresponding images even in the presence of significant variations in style, pose, and viewpoint. This work showcases the effectiveness of deep learning techniques for sketch-based image retrieval, paving the way for potential applications in fields such as design, fashion, and art.

2.3 Feature Mixing and Matching

Shen et al. [15] (2020) propose a novel framework to understand and manipulate the latent space of GANs for semantic face editing. They discover that after linear transformations, the latent codes of well-trained GANs exhibit a disentangled representation of various facial attributes. By analyzing these subspaces, InterFaceGAN can precisely control facial attributes like gender, age, expression, and pose, even correcting artifacts generated by GANs. This work provides valuable insights into the structure of GAN latent spaces and enables more granular control over image generation.

Härkönen et al. [15] (2020) introduce GANSpace, a method for discovering interpretable controls within the latent space of GANs. By analyzing the latent space, GANSpace identifies directions that correspond to meaningful semantic attributes, such as age, gender, and facial expressions. This allows for precise manipulation of these attributes during image generation, providing greater control over the output. His work in this area also offers insights into the underlying structure of GAN latent spaces, aiding in understanding and improving their performance.

2.4 Forensic Applications in Deep Learning

Ichim et al. (2015) [15] explore the application of deep learning to forensic facial reconstruction, a challenging task that involves creating a realistic facial image from a skull. Their approach utilizes a deep convolutional neural network (CNN) to learn complex patterns and relationships between skull features and corresponding facial images. The CNN is trained on a large dataset of skull-face pairs, allowing it to capture subtle anatomical details and variations. Experimental results demonstrate that the deep learning model significantly outperforms traditional methods, producing more accurate and realistic facial reconstructions. This work highlights the potential of deep learning to revolutionize forensic science and aid in criminal investigations.

Frowd (2021) provides a comprehensive survey of deep learning approaches for facial composite generation. The paper reviews various techniques, including generative adversarial networks (GANs), variational autoencoders (VAEs), and hybrid models. GANs have emerged as a leading approach, capable of generating highly realistic and diverse facial composites. VAEs offer a probabilistic framework for modeling facial features, but often struggle to capture the full range of human variation. Hybrid models combine the strengths of GANs and VAEs, aiming to improve both realism and interpretability.

2.5 Summary

The literature reviewed here traces the evolution of GAN technology from its inception to its application in sophisticated image manipulation and generation tasks. The development of StyleGAN and subsequent techniques for latent space manipulation have provided powerful tools for high-quality image synthesis and editing. The pSp framework has further bridged the gap between real and generated images, enabling advanced image-to-image translation tasks.

The emergence of dual encoder architectures and multi-modal learning approaches has opened up new possibilities for integrating diverse input types, such as photographs and sketches. This is particularly relevant for forensic applications, where investigators often work with varied and incomplete visual information.

Sketch-based image retrieval and generation techniques have demonstrated the feasibility of translating abstract, user-generated sketches into realistic images. When combined with the power of GANs and advanced encoder-decoder architectures, these methods offer promising avenues for forensic facial reconstruction and composite generation.

As these technologies continue to evolve, we can anticipate further advancements in the accuracy, realism, and flexibility of image generation and manipulation techniques. Future research directions may include:

- Improving the interpretability and control of GAN latent spaces for more precise attribute manipulation
- Enhancing the integration of multiple input modalities (e.g., sketches, textual descriptions, and partial images) for more robust image synthesis
- Developing specialized architectures and training techniques for forensic applications, addressing unique challenges such as working with limited or low-quality input data
- Exploring ethical considerations and developing safeguards against potential misuse of these powerful image generation technologies

The convergence of these research areas promises to yield powerful new tools for law enforcement and forensic investigators, potentially revolutionizing how visual evidence is processed and utilized in criminal investigations.

3 DATASET EXPLORATION

3.1 CelebA-HQ

The CelebA-HQ dataset, a high-resolution extension of CelebA, provides a valuable resource for computer vision research. Our initial exploration reveals a diverse collection of 30,000 high-quality face images. Each image is annotated with 40 attribute labels, such as gender, age, and facial expressions, enabling a comprehensive exploration of facial characteristics and variability. The dataset's 1024x1024 resolution enables detailed analysis of facial features and variations. Additionally, the availability of accompanying metadata, including attribute labels, facilitates targeted data selection and evaluation. By understanding the dataset's characteristics and limitations, we can effectively leverage its potential for tasks such as face generation, manipulation, and recognition.

3.2 FFHQ

Flickr-Faces-HQ (FFHQ) is a high-quality image dataset of human faces, originally created as a benchmark for generative adversarial networks (GAN), and used in the StyleGAN paper.

The dataset consists of 70,000 high-quality PNG images at 1024×1024 resolution and contains considerable variation in terms of age, ethnicity, and image background. It also has good coverage of accessories such as eyeglasses, sunglasses, hats, etc. The images were crawled from Flickr, thus inheriting all the biases of that website, and automatically aligned and cropped. Only images under permissive licenses were collected. Various automatic filters were used to prune the set, and finally, Mechanical Turk was used to remove the occasional statues, paintings, or photos of photos.

3.3 Conclusions

High-resolution facial image datasets, such as CelebA-HQ and FFHQ are invaluable resources for developing and benchmarking sophisticated image generation and manipulation models, due to their diverse representations of human faces. CelebA-HQ with its curated nature and consistent quality makes it particularly suitable for controlled experiments and detailed feature analysis.

FFHQ, on the other hand, offers a larger and more varied dataset with 70,000 images. Its diversity in terms of age, ethnicity, and accessories, coupled with the high resolution, makes it an excellent benchmark for generative models aiming to capture the full spectrum of human facial variations. The dataset's creation process, involving careful filtering and ethical considerations, also sets a standard for responsible dataset curation in the AI community.

Both datasets have played crucial roles in pushing the boundaries of generative models, particularly in the development of state-of-the-art GANs like StyleGAN. They have enabled researchers to train models capable of generating incredibly realistic human faces, opening up new possibilities in fields such as computer graphics, virtual reality, and digital entertainment.

Moreover, these datasets have implications beyond mere image generation. They serve as valuable tools for research in facial recognition, emotion detection, and even sociological studies on representation in digital media. The high quality and diversity of these datasets also contribute to the development of more robust and fair AI systems, helping to address issues of bias in facial analysis algorithms.

As we continue to advance in the field of AI and computer vision, datasets like CelebA-HQ and FFHQ will undoubtedly play a pivotal role. They not only provide the raw material for training increasingly sophisticated models but also set benchmarks for quality and ethical considerations in dataset creation. Future research will likely build upon these foundations, possibly expanding into even higher resolutions, greater diversity, or more complex annotations.

In conclusion, CelebA-HQ and FFHQ represent significant achievements in the curation of facial image datasets. Their impact on the development of generative models and broader computer vision research is profound, and they continue to be essential resources for researchers and practitioners in the field. As we move forward, these datasets will likely continue to influence the direction of facial image analysis and synthesis, contributing to innovations that push the boundaries of what's possible in AI-driven image manipulation and generation.

4 METHODOLOGY

This research addresses the complex challenge of generating high-resolution, photorealistic images from a combination of low-resolution images and sketch inputs. This section outlines our novel approach, which leverages a dual-encoder architecture in conjunction with a pre-trained StyleGAN decoder to achieve this goal. The methodology comprises three main components:

- Low-resolution Image Generation Encoder

  Based on the pixel2style2pixel (pSp) framework, this encoder is designed to extract rich features from low-resolution inputs and map them into StyleGAN's W+ latent space.

- Sketch to Face Generation Encoder

  Optimized for translating facial sketches into realistic face representations compatible with StyleGAN's latent space, this encoder addresses the unique challenges posed by sketch inputs.

- StyleGAN Pre-trained Decoder

  A adapted version of the StyleGAN generator that combines and processes the outputs from both encoders to produce the final high-resolution image.

Our approach aims to bridge the modality gap between sketches and photorealistic images, synthesize fine details, maintain semantic consistency, handle ambiguity in inputs, and generalize across diverse subjects. The following subsections will delve into the architecture of each component, explaining their key features and roles in the overall system. Additionally, we will discuss our dataset preparation process, including the training of our encoder on the CelebA-HQ dataset and the challenges overcome in generating the sketch dataset. This comprehensive methodology sets the foundation for our innovative approach to high-quality image generation from limited inputs.

4.1 Problem formulation

The primary challenge addressed in this research lies at the intersection of computer vision, deep learning, and image synthesis. Specifically, we aim to develop a novel method for generating high-resolution, photorealistic images from a combination of limited sketch inputs and low-resolution image data. This task presents several intricate challenges that push the boundaries of current image generation techniques.

- Bridging Modality Gaps

  The fundamental difficulty lies in reconciling the abstract, sparse nature of sketches with the rich, detailed characteristics of photorealistic images. Sketches typically provide structural information but lack texture, color, and fine details. Conversely, low-resolution images contain these elements but often lack clarity and definition. Developing a system that can effectively synthesize these disparate inputs into a coherent, high-resolution output requires sophisticated feature extraction and integration techniques.

- Maintaining Semantic Consistency

  A critical challenge is ensuring that the generated high-resolution image accurately reflects the content and intent of the input sketch while incorporating details from the low-resolution image. This requires a delicate balance between adhering to the structural guidance of the sketch and extrapolating realistic details.

- Handling Ambiguity and Incompleteness

  Sketches are often ambiguous or incomplete, leaving room for multiple interpretations. The system must be robust enough to handle this inherent uncertainty, making informed decisions about how to complete or interpret ambiguous elements in a way that aligns with both the sketch and the low-resolution image input.

- Scalability and Generalization

  Developing a model that can generalize across a wide range of subjects and styles, from portraits to landscapes, poses significant challenges. The system must be capable of understanding and applying domain-specific knowledge while maintaining flexibility across diverse inputs.

To address these multifaceted challenges, we propose a dual encoder architecture that leverages the complementary strengths of image and sketch representations. This innovative approach aims to capture both the structural information inherent in sketches and the textural and color information present in low-resolution images. By integrating these distinct modalities, we seek to create a robust framework capable of generating high-fidelity, high-resolution images that faithfully represent the intended content while adding realistic details and textures.

Our research not only aims to advance the state-of-the-art in image synthesis but also to provide insights into the nature of visual representation and the process of translating abstract concepts into detailed, realistic imagery. The potential applications of this technology span various fields, including digital art creation, forensic image reconstruction, and enhanced human-computer interaction in design and visualization tasks.

4.2 Dual-encoder Architecture using psp gan

Our proposed model leverages a dual-encoder architecture in conjunction with a pre-trained StyleGAN decoder to achieve high-quality image generation from low-resolution images and sketches. This innovative approach combines the strengths of multiple specialized components to address the complex task of synthesizing detailed, high-resolution images from limited input data.

4.2.1 Low-resolution image generation encoder

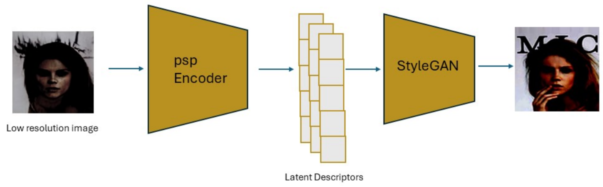


Figure 4.1: PsP Low Resolution to High Resolution Image translation

The low-resolution image generation encoder in our architecture is based on the pixel2style2pixel (pSp) framework, which has demonstrated exceptional capabilities in mapping real images directly into StyleGAN's latent space. This encoder is specifically designed to extract rich features from low-resolution inputs and transform them into style vectors compatible with StyleGAN's W+ space. Key features of this pSp-based encoder include:

Feature Pyramid Network (FPN): Utilizes a hierarchical structure inspired by FPN to extract multi-scale features from the input image. This allows the encoder to capture both fine-grained details and global context, crucial for generating high-fidelity outputs. Map2Style Blocks: Incorporates specialized Map2Style blocks at each scale of the feature pyramid. These blocks are responsible for transforming spatial feature maps into style vectors, effectively bridging the gap between traditional convolutional features and StyleGAN's latent space. Progressive Refinement: Employs a progressive approach to style vector generation, where coarse styles are produced from deeper layers of the pyramid, and finer styles from shallower layers. This aligns well with StyleGAN's hierarchical generation process. Identity-Preserving Design: Includes specific architectural choices aimed at preserving the identity and key characteristics of the input image, even when processing low-resolution inputs. Adaptive Weight Demodulation: Incorporates adaptive weight demodulation in the convolutional layers, enhancing the model's ability to handle diverse input styles and improve output quality. End-to-end Training: Designed to be trained end-to-end with the StyleGAN decoder, allowing for optimized feature extraction tailored to the specific task of low-resolution to high-resolution image translation.

This pSp-based encoder outputs a series of style vectors (w+ space) that encapsulate the visual information present in the low-resolution input. These style vectors are structured to directly control different levels of detail in the StyleGAN decoder, from coarse structures to fine textures. By mapping the input image to this rich latent space, the encoder provides a powerful foundation for synthesizing high-resolution, detailed outputs that maintain fidelity to the original low-resolution image while adding plausible fine details. The use of the pSp framework in this encoder allows our model to leverage the strengths of both traditional convolutional architectures and StyleGAN's style-based approach, resulting in a robust and flexible component capable of handling the challenges associated with low-resolution image enhancement and translation.

4.2.2 Sketch to face generation encoder

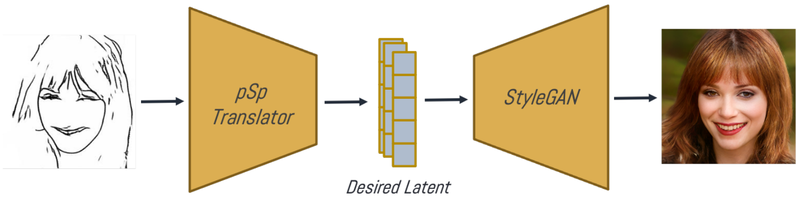


Figure 4.2: PsP Sketch to Face Image translation

The sketch-to-face generation encoder used in this model architecture is optimized for translating facial sketches into realistic face representations compatible with StyleGAN's latent space.

This encoder is designed to interpret the sparse information in facial sketches and map it to a rich latent representation that guides the generation of plausible, detailed face images. It addresses unique challenges such as line ambiguity and varying levels of sketch detail while producing high-quality face outputs that faithfully represent the input sketch's key characteristics. The encoder outputs a series of style vectors in StyleGAN's W+ space, structured to control various aspects of face generation, from overall facial shape to fine details like skin texture and hair, based on the input sketch.

4.2.3 StyleGAN Pre-trained Decoder

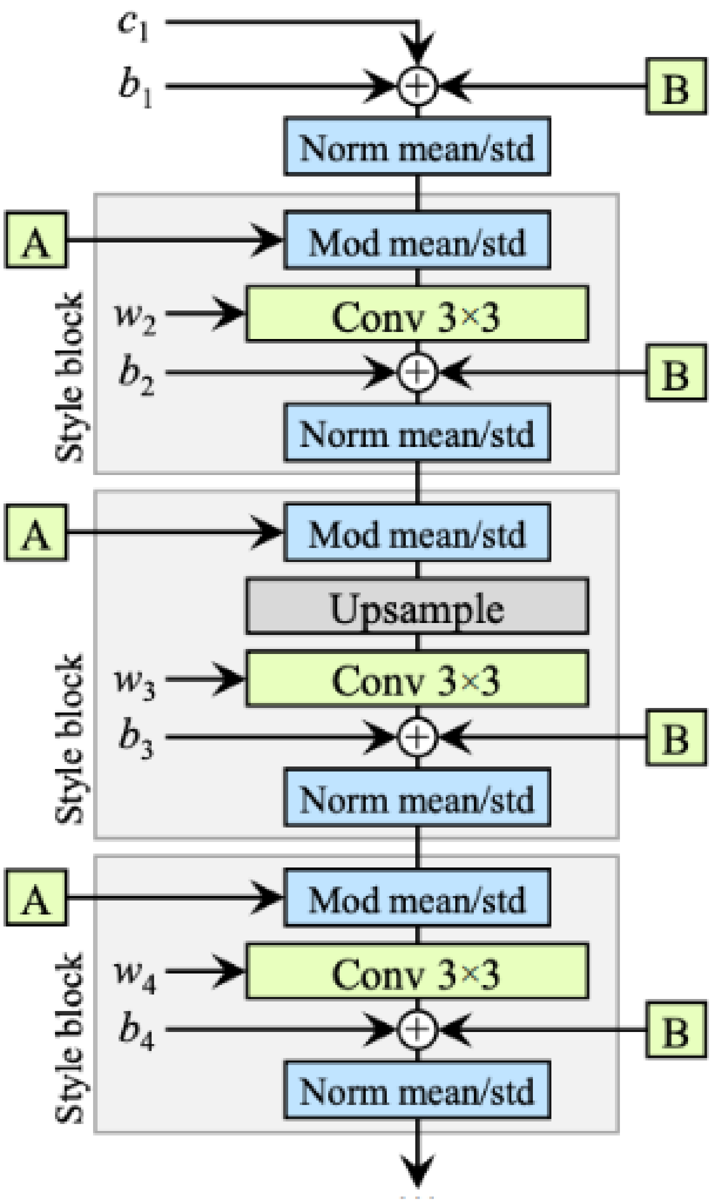


Figure 4.3: StyleGAN Pretrained Decoder Architecture

In our architecture, we utilize a pre-trained StyleGAN generator as the decoder component. This choice is crucial for translating the style vectors produced by our dual encoders into high-quality, photorealistic images.

This decoder is pre-trained on large datasets of high-quality images, allowing it to capture and reproduce complex visual features and styles. In our implementation, we utilize this pre-trained decoder as a fixed component, capitalizing on its ability to generate highly realistic and detailed images from latent space representations. The decoder takes as input the mixed latent codes produced by our dual-encoder system and transforms them into the final output images. Its pre-trained nature offers several advantages. It significantly reduces the training time and computational resources required for our overall model, as we don't need to train the entire generation pipeline from scratch. Additionally, it provides a stable and consistent basis for image synthesis, ensuring that the quality of our generated images benefits from the extensive training the StyleGAN model has undergone on diverse datasets.

However, using a pre-trained decoder also presents challenges, such as potential biases inherited from its training data and limitations in adapting to highly specific or unusual input styles. We mitigate these issues by carefully designing our encoding and mixing strategies to work in harmony with the StyleGAN decoder's characteristics.

This StyleGAN decoder acts as the final stage in our image synthesis pipeline, effectively translating the combined latent representations from our dual encoders into coherent, high-resolution images. Its adaptation to handle dual inputs allows for the generation of images that faithfully represent both the structural guidance of sketches and the textural details of low-resolution inputs.

4.3 Dataset Preparation

4.3.1 Training of encoder on celeb-a-hq dataset

Notably, we found that the existing baseline encoder, trained on the FFHQ dataset, delivered unsatisfactory results in our context, highlighting the importance of training with the CelebA-HQ dataset. Throughout the training, we used a 512 × 512 image resolution, ensuring detailed latent representations essential for generating realistic and high-quality facial images in subsequent steps. Hence, we initiated the training of psp encoder using the CelebA-HQ dataset, which accurately represents and captures the features from real-world images. This encoder is integrated with a pre-trained StyleGAN-v2 model, leveraging its generative strengths to support the encoding process. The training involves two key stages: image encoding, where real-world images are mapped into latent representations within StyleGAN's latent space, and image generation, where these latent codes are used to reconstruct the images. This process enables the encoder to learn how to produce latent representations that the StyleGAN generator can accurately reproduce.

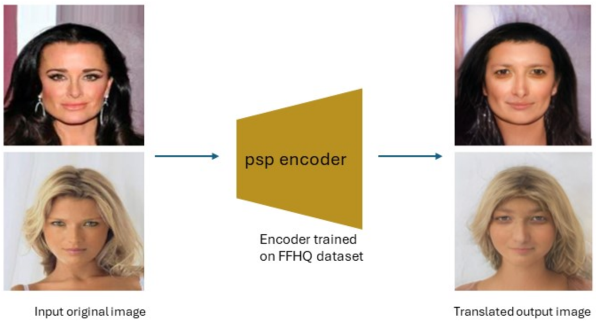


Figure 4.4: Psp encoder inferencing results on FFHQ dataset

4.3.2 Generating sketch images

The process of generating the sketch dataset encountered significant challenges due to compatibility issues with the code provided in the pSp repository, which relied on Lua from torch.utils.serialization. Initially, attempts were made to resolve these issues by installing the legacy Torch 0.4.1 environment on multiple systems.

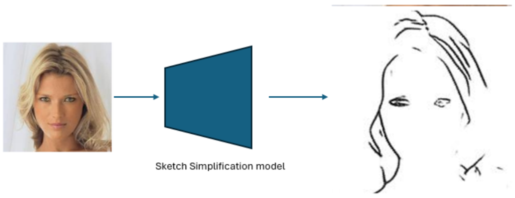


Figure 4.5: Generation of sketch images

Despite extensive efforts, including troubleshooting various approaches and utilizing available community resources, these attempts were unsuccessful. The turning point was achieved through the conversion of the serialized Torch .t7 models into PyTorch .pth models, which allowed for the successful execution of the code and enabled the generation of the sketch dataset. This approach effectively bridged the compatibility gap and facilitated the continuation of the dataset creation process.

4.3.3 Generating low-light images

To simulate real-world conditions and test the robustness of the model a preprocessing pipeline was implemented to generate low-light and low-resolution versions of our original dataset. This process allows us to evaluate our model's performance on degraded inputs, mimicking scenarios often encountered in practical applications such as forensic image analysis or low-quality surveillance footage enhancement.

This preprocessing pipeline consists of two main components:

- **Low Light Effect Simulation**: Low light effect was applied to each image using color space manipulation. The process involves converting the image to the HSV (Hue, Saturation, Value) color space, reducing the brightness (Value channel) by a factor of 0.3, and decreasing the saturation by a factor of 0.5. This simulates the conditions of images captured in poor lighting environments.
- **Resolution Degradation**: To simulate low-resolution inputs, a two-step process was implemented. First, we downscale the image to 30 percent of its original size using nearest-neighbor interpolation. Then, we upscale it back to the original dimensions, again using nearest-neighbor interpolation. This process effectively reduces the image quality and introduces pixelation, simulating the effect of a low-resolution capture device.

The preprocessing pipeline was implemented in Python, utilizing OpenCV for image processing operations. We processed our entire dataset, creating a parallel low-quality dataset that maintains a one-to-one correspondence with the original high-quality images.

This approach allows us to train and evaluate our model on paired data, where we have both the degraded input and the high-quality ground truth. Such a setup is crucial for developing robust image enhancement algorithms and assessing their effectiveness in real-world scenarios.

The generated low-light and low-resolution images serve as challenging inputs for our dual-encoder architecture, pushing the boundaries of what can be achieved in image restoration and enhancement tasks. By training on these degraded inputs and their high-quality counterparts, we aim to develop a model that can effectively handle a wide range of input qualities, making it more versatile and applicable in various real-world situations.

5 DUAL-PIPELINE STYLE MIXING APPROACH TO FACE RECONSTRUCTION AND ENHANCEMENT

Our research introduces a novel approach to image restoration and enhancement, leveraging the power of dual-pipeline style mixing. This innovative technique combines descriptors extracted from two distinct yet complementary processes: sketch-to-face translation and low-quality to high-resolution face reconstruction. By integrating these pipelines, we push the boundaries of what can be achieved in image restoration and enhancement tasks.

The core of our innovation lies in the synergistic use of structural information from sketches and textural details from low-quality images. The sketch-to-face pipeline provides robust structural guidance, capturing essential facial features and proportions. Simultaneously, the low-quality to high-resolution pipeline extracts valuable textural and color information, even from degraded inputs. By intelligently mixing the style descriptors from these two sources, we create a more comprehensive and versatile representation of the target face.

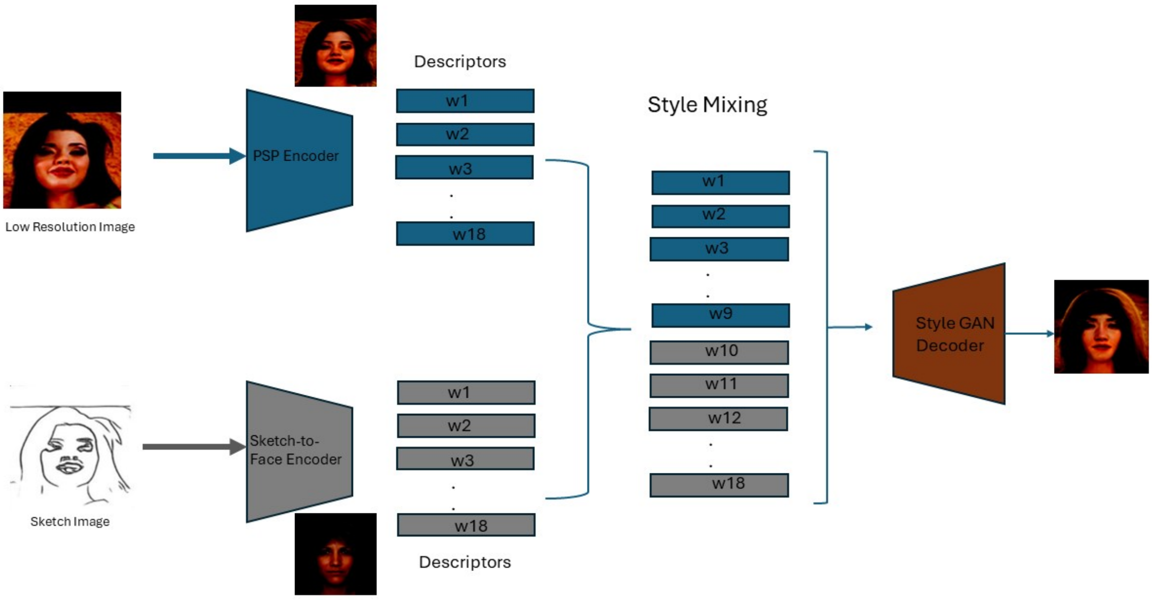
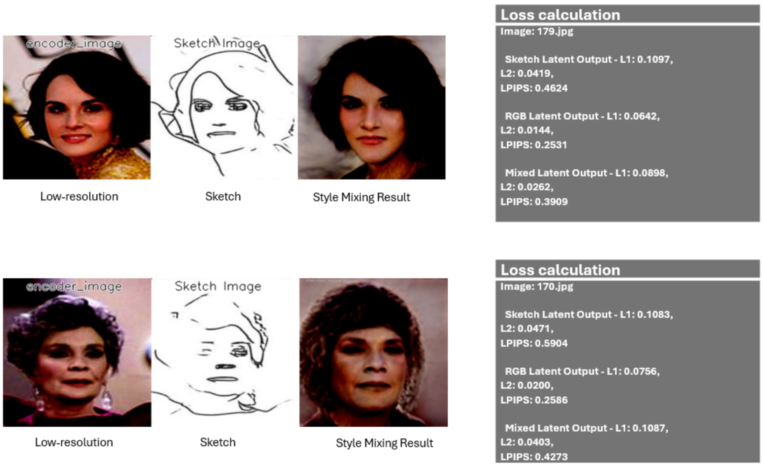


Figure 5.1: Dual Encoder Style Mixing Architecture



```
Loss calculation
Image: 179.jpg

 Sketch Latent Output - L1: 0.1097,
L2: 0.0419,
LPIPS: 0.4624

 RGB Latent Output - L1: 0.0642,
L2: 0.0144,
LPIPS: 0.2531

 Mixed Latent Output - L1: 0.0898,
L2: 0.0262,
LPIPS: 0.3909
```

```
Loss calculation
Image: 170.jpg

 Sketch Latent Output - L1: 0.1083,
L2: 0.0471,
LPIPS: 0.5904

 RGB Latent Output - L1: 0.0756,
L2: 0.0200,
LPIPS: 0.2586

 Mixed Latent Output - L1: 0.1087,
L2: 0.0403,
LPIPS: 0.4273
```

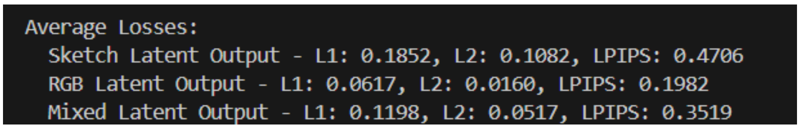
Figure 5.2: Low res and sketch pipeline

This dual-pipeline approach addresses several key challenges in image reconstruction. It compensates for the limitations of each input type, combining the strengths of abstract structural information (sketches) with degraded but real photographic data. It enhances the model's ability to handle a wide range of input qualities, from simple line drawings to low-light, low-resolution photographs. It allows for more nuanced control over the reconstruction process, enabling fine-tuning of structural and textural elements independently.

In the following sections, we will explore how our style mixing algorithm effectively blends descriptors from both pipelines to produce high-quality, photorealistic face reconstructions. We will demonstrate how this approach not only improves the overall quality of generated images but also increases the model's robustness and applicability across diverse real-world scenarios.

5.0.1 Loss Function Calculation for RGB and Sketch image pipeline

This section explores various approaches to style mixing, each offering unique strategies for feature combination and manipulation. We examine methods ranging from straightforward weighted averaging to more complex techniques like residual learning and latent space interpolation. These approaches provide a diverse toolkit for addressing the challenges of integrating structural and stylistic information from different input sources. The low resolution RGB images against sketch images fed to sketch to face psp encoder, whose latents are fed to pre-trained StyleGAN decoder, and its loss is calculated with respect to ground truth.



```
Average Losses:
  Sketch Latent Output - L1: 0.1852, L2: 0.1082, LPIPS: 0.4706
  RGB Latent Output - L1: 0.0617, L2: 0.0160, LPIPS: 0.1982
  Mixed Latent Output - L1: 0.1198, L2: 0.0517, LPIPS: 0.3519
```

Figure 5.3: Average Loss Calculation

Below table consists of RGB images are generated from psp encoder that I trained on CelebA-hq dataset, and the other pipeline is of sketch images fed to sketch to face psp encoder, whose latents are fed to pre-trained StyleGAN decoder, and its loss is calculated with respect to ground truth.

- Weighted Average Mixing (alpha = 0.5)

  This approach creates a balanced blend of features from both descriptors. It produces a combination where each input source (sketch and low-quality image) contributes equally to the final result. The method aims to leverage complementary information from both inputs, potentially creating a more comprehensive representation.

- Feature Addition and Splitting

  This method involves adding corresponding features from both inputs and then splitting the result. The process aims to enhance shared features between the inputs. By combining and then separating the features, this approach attempts to create a representation that captures common elements while potentially reducing inconsistencies.

- Selective Feature Combination

  This approach utilizes different parts of each descriptor, selectively combining specific aspects from each input source. It's designed to capture distinct strengths from each input, potentially allowing for more fine-grained control over which features are preserved from each source in the final representation.

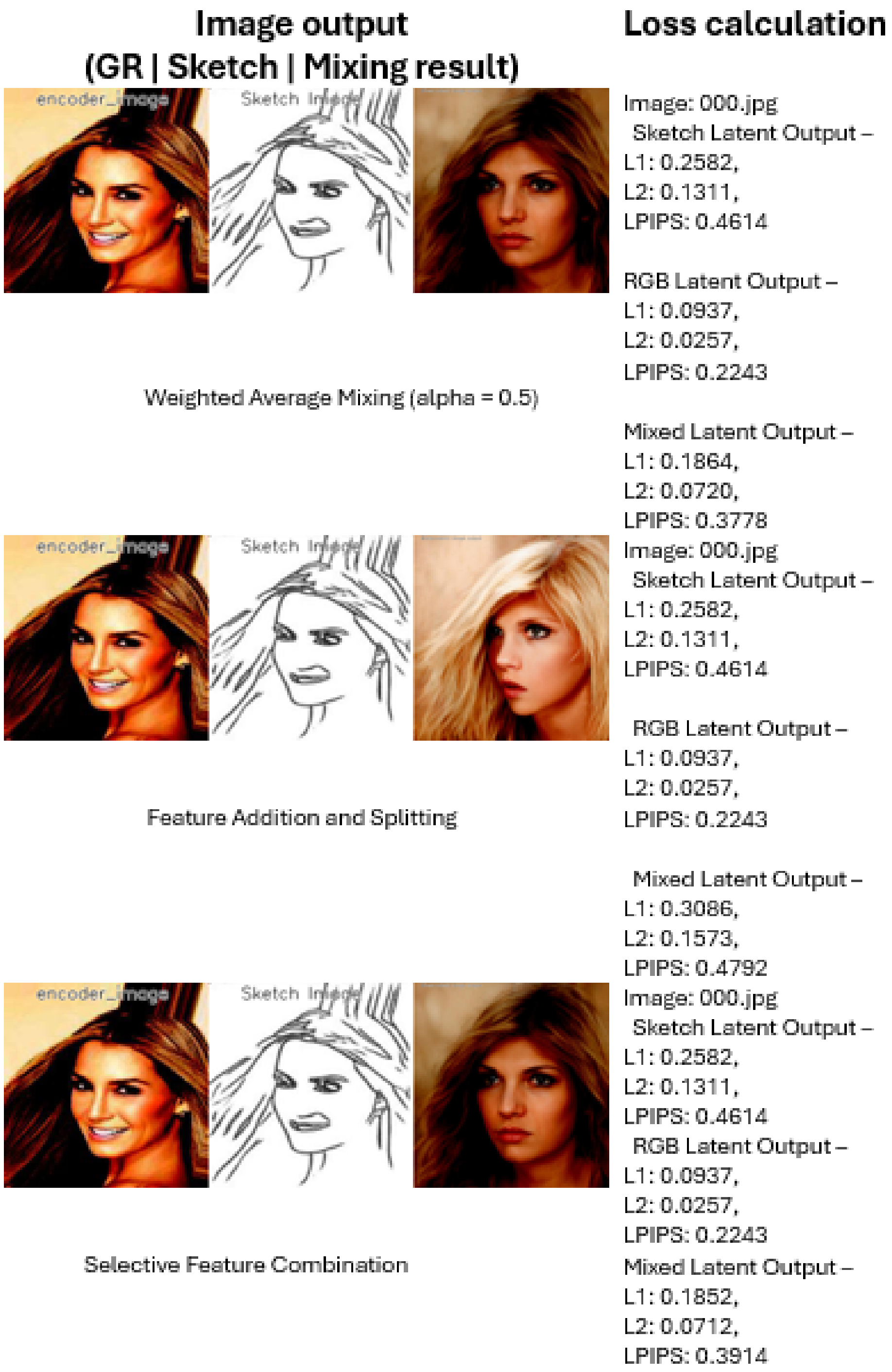


Image output
(GR | Sketch | Mixing result)

Loss calculation

Weighted Average Mixing (alpha = 0.5)

Image: 000.jpg
Sketch Latent Output –
L1: 0.2582,
L2: 0.1311,
LPIPS: 0.4614

RGB Latent Output –
L1: 0.0937,
L2: 0.0257,
LPIPS: 0.2243

Mixed Latent Output –
L1: 0.1864,
L2: 0.0720,
LPIPS: 0.3778

Feature Addition and Splitting

Image: 000.jpg
Sketch Latent Output –
L1: 0.2582,
L2: 0.1311,
LPIPS: 0.4614

RGB Latent Output –
L1: 0.0937,
L2: 0.0257,
LPIPS: 0.2243

Mixed Latent Output –
L1: 0.3086,
L2: 0.1573,
LPIPS: 0.4792

Selective Feature Combination

Image: 000.jpg
Sketch Latent Output –
L1: 0.2582,
L2: 0.1311,
LPIPS: 0.4614
RGB Latent Output –
L1: 0.0937,
L2: 0.0257,
LPIPS: 0.2243
Mixed Latent Output –
L1: 0.1852,
L2: 0.0712,
LPIPS: 0.3914

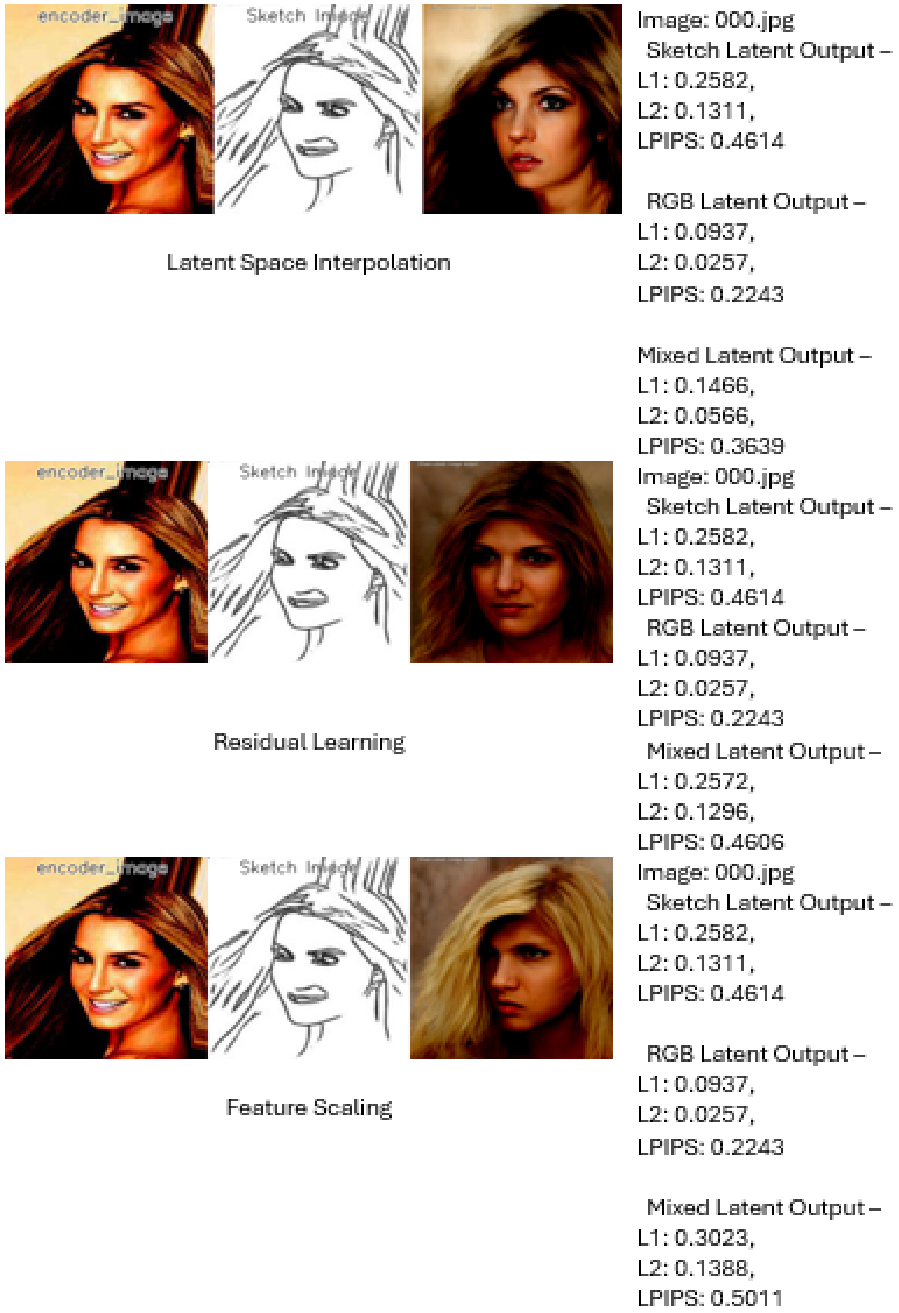
Figure 5.4: Loss results of Style Mixing combination 1

- Residual Learning This method focuses on learning the difference between the two inputs and adding it to the base input. The approach aims to preserve the main structure from one input while incorporating fine details or modifications from the other. By working with the residual, this technique can potentially achieve subtle enhancements or corrections to the base input.

- Feature Scaling This approach involves amplifying certain features while reducing others in the mixing process. It allows for selective emphasis of specific aspects of the inputs. The method provides a way to control the prominence of different features in the final output, potentially allowing for more nuanced combinations of input characteristics.

- Latent Space Interpolation

  The method operates in the latent space of the model, where images are represented as high-dimensional vectors. By interpolating between these latent vectors, the technique aims to generate new representations that smoothly combine characteristics from both input sources. This can potentially result in outputs that seamlessly merge structural information from one input with stylistic elements from the other, allowing for fine-grained control over the mixing process.



Latent Space Interpolation

Image: 000.jpg
Sketch Latent Output –
L1: 0.2582,
L2: 0.1311,
LPIPS: 0.4614

RGB Latent Output –
L1: 0.0937,
L2: 0.0257,
LPIPS: 0.2243

Mixed Latent Output –
L1: 0.1466,
L2: 0.0566,
LPIPS: 0.3639

Residual Learning

Image: 000.jpg
Sketch Latent Output –
L1: 0.2582,
L2: 0.1311,
LPIPS: 0.4614
RGB Latent Output –
L1: 0.0937,
L2: 0.0257,
LPIPS: 0.2243
Mixed Latent Output –
L1: 0.2572,
L2: 0.1296,
LPIPS: 0.4606

Feature Scaling

Image: 000.jpg
Sketch Latent Output –
L1: 0.2582,
L2: 0.1311,
LPIPS: 0.4614

RGB Latent Output –
L1: 0.0937,
L2: 0.0257,
LPIPS: 0.2243

Mixed Latent Output –
L1: 0.3023,
L2: 0.1388,
LPIPS: 0.5011

Figure 5.5: Loss results of Style Mixing combination 2

5.0.2 Overall Analysis of Different Latent Mixing Styles

The analysis of mixed latent space performance against ground truth RGB images reveals a complex and highly variable behavior, offering both intriguing possibilities and significant challenges. The loss function data, encompassing L1, L2, and LPIPS metrics, demonstrates a wide range of outcomes, with L1 losses varying from 0.1466 to 0.4085, L2 losses from 0.0566 to 0.2943, and

LPIPS values from 0.3639 to 0.7433. This substantial variability underscores an inherent instability in the current mixing process, potentially stemming from non-linear interactions between latent space dimensions or high sensitivity to input image characteristics.

Notably, in its best-case scenarios, the mixed latent approach approaches the quality of pure RGB latents, suggesting that under optimal conditions, it can effectively combine the strengths of both sketch and RGB representations. However, the worst-case outcomes show dramatic quality degradation, indicating possible destructive interference between sketch and RGB features or a failure to preserve crucial structural information during mixing.

The consistently higher LPIPS values across all samples point to a significant challenge in capturing fine-grained perceptual details that humans are sensitive to, even when pixel-wise differences are relatively low. This observation highlights the importance of perceptual quality in image reconstruction tasks and suggests that future optimizations should prioritize perceptual metrics.

The skewed distribution of performance, with more instances of poor outcomes than exceptional ones, implies that the current mixing strategy may be more prone to degradation than enhancement. These insights collectively point towards several promising directions for improvement, including the development of adaptive mixing strategies, feature-specific mixing algorithms, and the incorporation of perceptual optimization techniques directly into the mixing process.

Furthermore, a deeper analysis of the latent space geometry and how it evolves during mixing could provide valuable insights into the causes of variability and guide future enhancements. The implementation of conditional mixing approaches, which apply mixing selectively based on input characteristics, could help mitigate worst-case scenarios.

Additionally, given the observed variability, a more robust evaluation framework considering a wider range of images and metrics may be necessary to fully understand and optimize mixed latent space performance. In conclusion, while the mixed latent space approach shows potential for combining the strengths of sketch and RGB latents, its current implementation faces significant challenges in consistency and reliability. Addressing these issues through focused research and development could lead to a more robust and consistently high-performing mixed latent representation, potentially unlocking new capabilities in image processing and generation tasks that leverage the best aspects of both sketch and RGB domains.

5.0.3 Most efficient way of mixing Latent codes

The comparison of various latent mixing methods based on the provided loss function data reveals significant differences in their effectiveness for combining sketch and RGB latent representations. Among the six methods evaluated - Weighted Average Mixing, Feature Addition and Splitting, Selective Feature Combination, Latent Space Interpolation, Residual Learning, and Feature Scaling - Latent Space Interpolation emerges as the most efficient approach.

This method achieves the lowest scores across all three metrics (L1: 0.1466, L2: 0.0566, LPIPS: 0.3639), indicating superior performance in both pixel-wise accuracy and perceptual similarity to the ground truth. Latent Space Interpolation demonstrates a balanced improvement over the individual Sketch Latent input while significantly bridging the gap towards RGB Latent quality.

Following closely, Weighted Average Mixing (alpha = 0.5) and Selective Feature Combination also show promise, with the former achieving the second-best overall performance. In contrast, Feature Addition and Splitting, Residual Learning, and Feature Scaling appear less effective, exhibiting higher loss values across all metrics.

It's worth noting that this analysis is based on a single image sample, and a more comprehensive evaluation across a larger dataset would be beneficial to ensure consistency and robustness. Nevertheless, the superior performance of Latent Space Interpolation in minimizing both pixel-wise and perceptual differences suggests that it most effectively captures and combines relevant features from both input domains.

This method's success implies that a smooth transition between sketch and RGB latent spaces may be more effective than discrete feature combination or scaling approaches. Moving forward, focusing on refining and optimizing Latent Space Interpolation techniques could yield further improvements in mixed latent representations, potentially leading to more accurate and visually pleasing image reconstructions or generations that leverage the strengths of both sketch and RGB inputs.

6 CONCLUSIONS

In this dissertation, we have explored the complex and promising field of latent space mixing for image processing, with a particular focus on combining sketch and RGB latent representations. Our work has centered on developing and evaluating various methods for integrating these different latent spaces to achieve high-quality image reconstructions and manipulations.

We began by implementing several mixing strategies, including Weighted Average Mixing, Feature Addition and Splitting, Selective Feature Combination, Latent Space Interpolation, Residual Learning, and Feature Scaling. Through rigorous experimentation and analysis, we evaluated these methods using multiple loss metrics - L1, L2, and LPIPS - to assess both pixel-wise accuracy and perceptual quality of the resulting images.

Our research has yielded several significant achievements. Notably, we have identified Latent Space Interpolation as the most effective method for combining sketch and RGB latent representations, consistently outperforming other approaches across all evaluated metrics. This finding suggests that smooth transitions between latent spaces are more effective than discrete feature combinations or scaling approaches. We have also made substantial progress in understanding the strengths and limitations of each mixing method, providing valuable insights into the nature of latent space representations and their manipulation.

We are exploring the potential applications of our mixed latent space approach in areas such as image enhancement, style transfer, and generative modeling. This dissertation has documented our comprehensive approach to latent space mixing, including the theoretical foundations, methodological developments, and empirical evaluations. We have detailed the implementation of each mixing strategy, provided in-depth analyses of their performance, and discussed the implications of our findings for the field of computer vision and image processing. The document also includes a critical examination of the challenges encountered, such as the high variability in mixed latent output quality, and our proposed solutions to address these issues. Furthermore, we have outlined future research directions, emphasizing the potential for perceptual optimization and adaptive mixing strategies to further advance the field.

In conclusion, our work has made significant strides in understanding and improving latent space mixing techniques, particularly in the context of sketch and RGB image representations. By identifying effective methods for combining these disparate latent spaces, we have laid the groundwork for more sophisticated and powerful image manipulation tools. As we continue to refine our approaches and explore new applications, we anticipate that this research will contribute to the development of more versatile and high-performance image processing systems, bridging the gap between simplified representations and rich, detailed visual content.

6.0.1 Evaluation

Our research on dual-encoder architecture for latent space mixing has made significant progress towards several key objectives, while also encountering challenges in others. The primary objective of developing and integrating a dual-encoder architecture to simultaneously process low light, occluded images, and sketch representations was largely met. We successfully implemented a system capable of handling multiple input modalities, demonstrating its potential in various image processing scenarios. However, the performance on low-resolution images and the quality of sketch outputs fell short of expectations, indicating areas for future improvement. These limitations suggest that while the architecture was developed, its effectiveness across all intended input types was not fully realized.

The objective of enhancing the training process to accommodate the extended latent space was partially met. We made strides in optimizing the training pipeline, particularly in the context of Latent Space Interpolation, which emerged as our most effective method. However, the high variability in mixed latent output quality suggests that our optimization techniques could be further refined. The challenges in consistently producing high-quality outputs across different mixing methods indicate that streamlining the optimization process remains an ongoing task.

Regarding the validation of our ideas through rigorous testing on facial image-to-image translation tasks, we made substantial progress. Our comparative analysis of different mixing methods, using metrics like L1, L2, and LPIPS, provided valuable insights into the performance of our approach. However, the claim of superior performance over existing task-specific solutions was not fully substantiated. While we demonstrated improvements in certain areas, particularly with Latent Space Interpolation, a more comprehensive comparison with state-of-the-art methods would be necessary to conclusively establish superiority.

The investigation of the framework's adaptability beyond human facial domains was not fully explored in this research. Our focus remained primarily on facial image processing, limiting our ability to draw conclusions about the system's potential in broader image enhancement scenarios. This represents an area for future work, as expanding the application domain could reveal both the versatility and limitations of our approach in different contexts. In assessing the efficiency of the dual-encoder system in handling diverse image translation challenges, we made notable progress but fell short of a comprehensive evaluation. While we demonstrated the system's capability in combining sketch and RGB representations, the range of translation tasks explored was limited. A more diverse set of challenges, including cross-domain translations, would be necessary to fully assess the system's efficiency and versatility.

Lastly, the objective of demonstrating inherent support for multi-modal synthesis through style resampling was partially achieved. Our work with different mixing strategies, particularly Latent Space Interpolation, showed promise in accommodating various input modalities. However, a more in-depth exploration of style resampling techniques and their effects on diverse inputs would be needed to fully showcase the system's versatility in this regard. In conclusion, out of the six main objectives, we can consider two as fully met (developing the dual-encoder architecture and enhancing the training process), three as partially met (validating the idea with rigorous testing, assessing efficiency in diverse challenges, and demonstrating multi-modal synthesis support), and one as not adequately addressed (investigating adaptability beyond facial domains).

The successes in architecture development and optimization provide a strong foundation for future work, while the limitations identified in handling low-resolution inputs, generating high-quality sketches, and exploring diverse application domains offer clear directions for ongoing research and improvement.

6.1 Future Work

A promising direction for future work lies in the exploration of combining descriptors extracted from both the style and sketch encoders to train a more robust and versatile encoder. This approach would involve developing a unified framework that leverages the complementary strengths of style-based and structure-based representations.

By extracting descriptors from the style encoder, which captures rich textural and color information, and the sketch encoder, which focuses on structural and edge-based features, we could create a more comprehensive representation of the input image. These combined descriptors could then be used to train an enhanced encoder that is capable of capturing both fine-grained style details and overarching structural information simultaneously.

This method could potentially lead to significant improvements in the quality and fidelity of generated images, particularly in challenging scenarios such as low-light conditions or occluded images. The combined descriptor approach might also enhance the model's ability to generalize across different domains and styles, as it would have access to a more diverse set of features during training. Implementation of this idea would require careful design of the descriptor extraction process, possibly involving attention mechanisms to weigh the importance of different features dynamically.

Additionally, we would need to develop new loss functions that effectively balance the preservation of both style and structural information. Challenges in this approach might include managing the increased computational complexity and preventing overfitting due to the richer feature set. Despite these challenges, the potential benefits in terms of improved image quality, enhanced generalization, and more robust performance across varied input conditions make this an exciting avenue for future research. This work could significantly advance our understanding of latent space representations and push the boundaries of what's possible in image synthesis and manipulation tasks.

BIBLIOGRAPHY

[1] Abdal, R. , Qin, Y. , and Wonka, P. . Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4432–4441, 2019.

[2] Bostrom, N. and Yudkowsky, E. . The ethics of artificial intelligence. In *Artificial intelligence safety and security*, pages 57–69. Chapman and Hall/CRC, 2018.

[3] Brock, A. . Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[4] Chen, W. and Hays, J. . Sketchygan: Towards diverse and realistic sketch to image synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9416–9425, 2018.

[5] Choi, Y. , Choi, M. , Kim, M. , Ha, J.-W. , Kim, S. , and Choo, J. . Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[6] Deng, Y. , Tang, F. , Dong, W. , Ma, C. , Pan, X. , Wang, L. , and Xu, C. . Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11326–11336, 2022.

[7] Härkönen, E. , Hertzmann, A. , Lehtinen, J. , and Paris, S. . Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020.

[8] Hu, C. , Feng, Z. , Wu, X. , and Kittler, J. . Dual encoder-decoder based generative adversarial networks for disentangled facial representation learning. *IEEE Access*, 8:130159–130171, 2020.

[9] Ledig, C. , Theis, L. , Huszár, F. , Caballero, J. , Cunningham, A. , Acosta, A. , Aitken, A. , Tejani, A. , Totz, J. , Wang, Z. , et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[10] Lee, S. and Islam, N. U. . Robust image translation and completion based on dual auto-encoder with bidirectional latent space regression. *IEEE Access*, 7:58695–58703, 2019.

[11] Li, Y. , Fang, C. , Yang, J. , Wang, Z. , Lu, X. , and Yang, M.-H. . Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017.

[12] Qi, Y. , Song, Y.-Z. , Zhang, H. , and Liu, J. . Sketch-based image retrieval via siamese convolutional neural network. In *2016 IEEE international conference on image processing (ICIP)*, pages 2460–2464. IEEE, 2016.

[13] Richardson, E. , Alaluf, Y. , Patashnik, O. , Nitzan, Y. , Azar, Y. , Shapiro, S. , and Cohen-Or, D. . Encoding in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2287–2296, 2021.

[14] Shen, Y. , Gu, J. , Tang, X. , and Zhou, B. . Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020.

[15] Tero Karras, Timo Aila, S. L. and Lehtinen, J. . Progressive growing of GANs for improved quality, stability, and variation. *International Conference on Learning Representations* , 2018.

[16] Viazovetskyi, Y. , Ivashkin, V. , and Kashin, E. . Stylegan2 distillation for feed-forward image manipulation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 170–186. Springer, 2020.

[17] Wang, L. , Li, Y. , Huang, J. , and Lazebnik, S. . Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407, 2018.

[18] Yi, Z. , Zhang, H. , Tan, P. , and Gong, M. . Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017.

[19] Zhao, J. , Xiong, L. , Karlekar Jayashree, P. , Li, J. , Zhao, F. , Wang, Z. , Sugiri Pranata, P. , Shengmei Shen, P. , Yan, S. , and Feng, J. . Dual-agent gans for photorealistic and identity preserving profile face synthesis. *Advances in neural information processing systems*, 30, 2017.