Efficient Identity-preserving Face Swapping in Scenic Sketches

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering
by Research

by

Ankith Varun J
2020111019
ankith.varun@research.iiit.ac.in

Advisor: Dr. Anoop Namboodiri

International Institute of Information Technology Hyderabad
500 032, India

October 2025

International Institute of Information Technology Hyderabad
Hyderabad, India

CERTIFICATE

This is to certify that work presented in this thesis proposal titled ***Efficient Identity-preserving Face Swapping in Scenic Sketches*** by *Ankith Varun J* has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Dr. Anoop Namboodiri

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. Anoop Namboodiri, for their unwavering support, insightful guidance, and encouragement throughout the duration of my research journey. I sincerely thank him for granting me the opportunity to work at the Center for Visual Information Technology (CVIT) and Biometrics and Secure Identity Lab (BaSIL) and for his invaluable expertise and feedback which were instrumental in shaping this work.

I would also like to thank my mentors, Kushal Kumar Jain and Tathagato Roy, for their continual help, advice, and inspiration. Their knowledge and experience have greatly enriched my academic journey.

Special thanks are due to my fellow researchers and collaborators, including Shreya Bollimuntha, Gaurav Bansal and Karuna K Chandra, for their contributions to the research process and their valuable input during the writing phase. Working alongside them was both intellectually stimulating and personally rewarding. Family and Friends

I am deeply grateful to my family for their unconditional love, patience, and support. To my parents, Jayaram G A and Vimala K P, thank you for believing in me and encouraging me to pursue my dreams. To my siblings, Vibhashree J and Varsha M, thank you for your unwavering support and for always being there when I needed guidance or a listening ear.

I would also like to express my gratitude to my extended family members who have been pillars of support throughout this journey: Pramila K P, Ravi Kumar K P, P R Leela, L K Panduranga, Gorur Anathanarayan and Premaleela N.

I am fortunate to have had the support of wonderful friends throughout different phases of my life. To my childhood friends Prakul Pandita, Aarya Sharma, Anoushka Asrani, Sourav M S and Kaustubh Chaturvedi, thank you for your lifelong friendship and for being there from the very beginning.

To my college friends who were by my side throughout this journey: Rohan Chowdary, Joel Alex, Siddharth Mavani, Aum Khatlawala, Khush Patel, Ayush Agrawal, Shubh Agarwal, Swetha Vipparla, Karuna K Chandra, Shreya Bollimuntha, Gaurav Bhole, Abraham Paul and Sumit Kumar, I am profoundly grateful for your unwavering support, late-night brainstorming sessions, and the deep bonds we've formed. I would also like to thank Harshit Gupta, Tathagato Roy, Kushal Jain, Likhith Asapu, Rajarshi Ray, Pranjal Thapliyal, Priyanshul Govil, Praneetha Gokul, Gokul Raj, Ruhul Ameen, Janardan S, Sukhjinder Kumar, Ben Paul, Aayush Bhandari, Amal Sunny, Tejasvi Chebrolu, Pratykash Gautam and Aneesh Chavan for the collaborative projects, stimulating discussions, and memorable moments that enriched my academic experience.

I would like to extend my heartfelt appreciation to my football team, who provided me with the perfect balance of physical activity and camaraderie during the intense periods of research. Our training sessions and matches served as welcome breaks from academic pressures and reminded me of the importance of teamwork both on and off the field.

I am especially grateful to our team captain, Abraham Paul, whose leadership and motivation pushed us to achieve our best. Special thanks to our coach, Revanth Nethala, for their tactical wisdom and patience.

To my teammates: Gaurav Bhole, Sumit Kumar, Joel Alex, Rohan Chowdary, Shreyash Jain, Vayur Shanbagh, Rohit Reddy, Vaibhaw Bhaiya, Kashif Mohammad, Aniket Gupta, Annamalai Senthil, Sajiv Singh, Anurag Dubey, Rohan Sridhar, Atidipt Ashnin, Sriram Alluri, Amey Bangera, Aayush Khatanhar, Ishaan Kumar, Arnav Agnihotri, Soham Acharya, Arjun Dingankar and Karthik Sundaram – thank you for the unforgettable moments, the victories we celebrated together, and the losses we overcame as a team. Our post-match discussions and team dinners offered me perspectives that often found their way into my research thinking.

The discipline, perseverance, and collaborative spirit I learned on the football field have been invaluable in my academic pursuits, and I am grateful to have been part of such an exceptional team.

This thesis would not have been possible without the collective support of all these wonderful individuals who have enriched my life and contributed to my growth both as a researcher and as a person.

Abstract

We present an efficient framework for identity-preserving face swapping into scenic portrait sketches. First, we analyze latent identity representations via a StyleGAN-based face cartoonization pipeline. Our analysis reveals domain-dependent shifts in StyleGAN's identity encodings when mapping photos to sketches, underscoring the need for cross-domain consistency. Second, we introduce Portrait Sketching StyleGAN (PS-StyleGAN), a novel GAN architecture with attentive affine-transform blocks. These blocks learn to modulate a pretrained StyleGAN's style codes by joint content-and-style attention, thereby learning style-specific identity transformations. PS-StyleGAN requires only a few photo-sketch pairs (and short training) to learn each style, making it broadly adaptable. Thus, PS-StyleGAN produces editable, expressive portrait sketches that faithfully preserve the input identity. Third, we extend diffusion-based generative modeling by adapting InstantID for sketch synthesis. Our diffusion module integrates strong semantic identity embeddings and landmark guidance (inspired by InstantID's IdentityNet) to steer high-fidelity portrait generation. This yields personalized sketches with markedly improved identity fidelity while requiring only zero-shot (tuning-free) personalization. Finally, motivated by recent diffusion-based face-swapping advances, we build a real-time end-to-end pipeline. By encoding facial identity and pose as conditioning inputs and optimizing the diffusion process for speed, our system can seamlessly embed user faces into tourist-style sketches on the fly. Extensive evaluations on standard benchmarks confirm that our methods significantly enhance identity preservation and visual quality compared to prior approaches, advancing the state of the art in generative portrait stylization and face swapping.

Contents

List of Figures

List of Tables

List of Related Publications

[P1] Kushal Kumar Jain, Ankith Varun, and Anoop Namboodiri, **"PS-StyleGAN: Illustrative Portrait Sketching using Attention-Based Style Adaptation"**, in proceedings of *27th International Conference on Pattern Recognition, Kolkata*, 2024.

[P2] Kushal Kumar Jain, Ankith Varun, and Anoop Namboodiri, **"Face Cartoonisation For Various Poses Using StyleGAN"**, preprint at arXiv, 2023.

Chapter 1

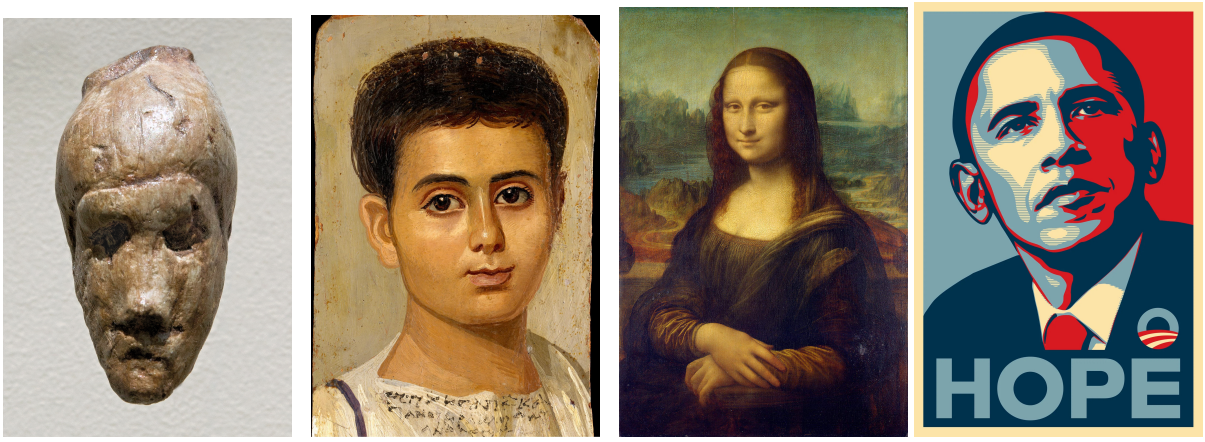
Introduction

Facial identity has a profound significance in human perception and social interaction. In biometric recognition systems, facial identity is classified as a physiological characteristic - a measurable biological trait that uniquely identifies an individual [1]. The combination of facial geometry, texture, and proportions creates a unique signature that the human brain is remarkably adept at processing and recognizing. While biometric approaches focus primarily on permanent physiological features, the concept of facial identity extends beyond this limited framework. Visual artforms throughout history reveal a more nuanced understanding, as they capture the stillness of a moment in an individual's life rather than just their permanent characteristics. This expanded conception of facial identity encompasses both permanent traits and impermanent features such as the frame of the face, facial hair, and other transient attributes that contribute to one's visual identity. From prehistoric cave paintings to Renaissance portraits and modern digital imagery, artists have documented this comprehensive view of facial identity through various mediums, demonstrating the timeless value of portraiture in human culture; see Fig. 1.1.

Portraiture as an art form inherently explores identity through visual representation. Beyond mere physical resemblance, portraits capture personality, social status, emotional states, and cultural contexts [2]. Artists make deliberate choices about style, composition, and technique to convey specific aspects of identity, resulting in works that transcend simple representation to become interpretative expressions of the subject's essence.

The case of artist Chuck Close provides a fascinating perspective on identity encoding in portraiture. Despite suffering from prosopagnosia (face blindness)—a neurological condition that impaired his ability to recognize faces in three-dimensional space - Close created remarkable photorealistic portraits [3]. His work demonstrates how two-dimensional artistic representation can systematically encode identity even when the artist's own facial recognition systems are compromised. Close developed a methodical grid-based approach to portraiture that allowed him to translate facial features into visual art with extraordinary precision despite his perceptual limitations.

From a neurological standpoint, facial identity recognition is a specialized cognitive function with dedicated neural pathways. Research indicates that the fusiform face area (FFA) in the temporal lobe plays a crucial role in facial recognition processes [4]. This dedicated neural architecture underscores



(a) An ivory carving found in Dolní Věstonice dated 24,000 BCE

(b) Egyptian funeral portrait for Eutyches of about 100 to 150 AD

(c) Oil painting of Mona Lisa by Leonardo Da Vinci during the 16th century

(d) Digital poster of Barack Obama by American artist Shepard Fairey in 2008

Figure 1.1: (a) The statuette made from mammoth ivory represents one of the earliest attempts to capture human likeness of a face. It can be recognized as a portrait because of its distinct individual features—such as an asymmetrical eye and a chin dimple—revealing it as the depiction of a real, living woman. (b) the Egyptian funeral portrait of Eutyches (100-150 AD) shows the evolution of identity preservation in ancient civilizations. (c) Leonardo Da Vinci's Mona Lisa from the 16th century demonstrates the Renaissance mastery of capturing subtle facial characteristics. (d) Shepard Fairey's digital poster of Barack Obama from 2008 illustrates how facial identity continues to be powerful in contemporary digital art.

the evolutionary importance of facial identity recognition for human social interaction and survival. The brain processes faces holistically rather than as collections of individual features, enabling rapid and accurate identification even under varying conditions of lighting, angle, and expression.

Across different visual media, identity representation varies significantly due to the inherent characteristics of each medium. In photography, identity is captured with high fidelity to physical appearance. In painting, the artist's interpretation and style influence the representation. In caricature, identity is expressed through the exaggeration of distinctive features. Each medium creates unique perceptual challenges and opportunities for identity preservation [5].

In the specific domain of sketches and line art, human perception operates under distinct constraints. As Hertzmann [6] explains, line drawings represent a significant abstraction from natural images, yet humans interpret them with remarkable ease. Lines in sketches often correspond to object boundaries, depth discontinuities, and surface orientation changes, forming a perceptual shorthand that the human visual system readily interprets. This abstraction poses unique challenges for computational systems that attempt to preserve identity in sketch-based representations.

This brings us to the core challenge addressed in this thesis: face swapping as a disentanglement problem. To successfully swap faces across domains, we must separate facial identity from other semantic attributes such as expression, pose, lighting, and style. This disentanglement becomes particularly

complex when transitioning between photorealistic and non-photorealistic domains, where the visual cues that signify identity may be represented in fundamentally different ways. Automating this process for a large-scale system requires sophisticated approaches that can maintain identity consistency while adapting to the stylistic requirements of the target domain.

Face swapping in non-photorealistic domains presents a complex challenge that can be decomposed into three interconnected problems:

1. **Identity Representation**: This involves extracting and encoding the distinctive characteristics that constitute an individual's facial identity. The challenge lies in creating representations that capture identity-specific features while remaining robust across different poses, expressions, and lighting conditions. In non-photorealistic domains such as sketches, the identity representation must adapt to the stylistic constraints of the medium.

2. **Face Reenactment**: This component focuses on transferring expressions, pose, and other non-identity attributes from a source face to a target identity. The system must preserve the target's identity while accurately reproducing the source's dynamic attributes, creating a natural-looking synthesis.

3. **Style Transfer**: This entails adapting the facial representation to conform to the stylistic characteristics of the target domain. For sketch-based face swapping, this means translating photorealistic facial details into line-based representations that follow the conventions of the sketch medium.

A comprehensive face swapping system must seamlessly integrate these three tasks, maintaining identity consistency while adapting to the stylistic requirements of the target domain and preserving the expressive characteristics of the source face.

The applications of such technology are diverse and impactful. In entertainment, it enables the creation of personalized content where users can insert themselves into various media. In education, it can facilitate engaging historical or scientific visualizations. In virtual reality, it supports enhanced avatar customization for more immersive experiences.

A particularly compelling application—and the focus of this thesis—is in tourist settings where visitors can obtain personalized scenic sketches featuring their own faces. This real-time face swapping system would capture a tourist's facial identity and seamlessly integrate it into artistic renderings of landmarks or landscapes, creating unique, personalized souvenirs. Such a system must operate efficiently with minimal latency to provide a smooth user experience in high-traffic tourist locations.

The primary objective of this thesis is to develop and analyze methods for performing cross-domain face swapping, with a particular focus on generating sketch-based representations that preserve facial identity. Furthermore, we aim to design an optimized system that satisfies real-world performance requirements for practical applications.

Our major contributions are as follows:

- We modify a face re-enactment pipeline to generate faces in the 'cartoon' domain. Through this work, we demonstrate that identity distributions shift significantly when changing domains, revealing a fundamental correlation between identity representation and style domain. This finding has important implications for cross-domain identity preservation techniques, suggesting that identity encodings must be adapted to account for domain-specific characteristics. (Chapter 3)

- We introduce PS-StyleGAN for portrait sketch generation, which directly addresses the domain-dependency of identity representation through novel style modulation wrappers. This approach explicitly models the correlation between identity features and artistic style to maintain identity consistency across domains. (Chapter 4)

- We explore and extend InstantID, a diffusion-based approach that generates superior results. We adapt this method specifically for the sketch domain using personalized styles, enabling high-quality portrait sketch generation with preserved identity features. (Chapter 5)

- We design an end-to-end system built on our modified InstantID pipeline, incorporating both qualitative and quantitative optimizations to achieve near real-time performance. This system balances computational efficiency with output quality, making it suitable for practical applications in tourist settings where responsiveness is crucial. (Chapter 6)

This thesis is organized into seven chapters. Following this introduction, Chapter 2 reviews the state-of-the-art in facial identity representation, face reenactment, style transfer, and face swapping, with particular attention to cross-domain applications. Chapter 3 explores domain-specific identity representation through a modified face re-enactment pipeline with cartoonization capabilities. Chapter 4 presents a GAN-based approach with PS-StyleGAN featuring attention-based style adaptation, while Chapter 5 a diffusion-based approach adapting InstantID for the sketch domain using DreamBooth, Textual Inversion and LoRA techniques. Chapter 6 develops an end-to-end system for real-time face swapping in sketch domains, with detailed analysis of qualitative and quantitative optimizations. Finally, Chapter 7 summarizes our findings, discusses limitations, and suggests directions for future research.

Chapter 2

Related Works

2.1 Identity Representation

High-fidelity face generation requires robust identity encodings. Early approaches used deep face recognition models [7, 8] to extract identity features, but these were trained for classification rather than generation. Seminal works such as CosFace [9] and ArcFace [10] introduced large-margin losses that greatly improved identity discrimination by increasing intra-class compactness and inter-class separability. Recent works by Zheng et al. [11] introduce FaRL, which enhances facial representation by refining the pretrained CLIP architecture [12]. Their approach develops more expressive facial token representations through fine-tuning on extensive datasets containing face-text pairs. These embeddings have been reused as identity encodings in generative models in the form of loss functions and conditional inputs. However, they remain optimized for single-view recognition, losing fine identity details across poses and expressions. Recognizing this limitation, recent works design identity representations specifically for generation. For example, Qian et al. [13] propose Omni-ID, a generative identity encoder that consolidates multiple images of a person into a single feature. Omni-ID is trained with a few-to-many reconstruction objective, so it learns holistic identity features across varied expressions and poses. This approach outperforms conventional discriminative encoders in preserving identity in new poses and styles but its integration into existing frameworks incurs huge training costs. In summary, face generation pipelines typically rely on pretrained identity encoders (ArcFace) over task-specific variants (like Omni-ID) to preserve identity when swapping and stylizing faces.

2.2 Style Transfer

Artistic style transfer has evolved from Non-Photorealistic Rendering (NPR) and classical neural methods to modern GAN and diffusion approaches. NPR methods [14–18] rely on ground truth geometry, which is noisy near detailed parts of the face like eyes, nose and lips resulting in diminished quality. Gatys et al. [19] pioneered neural style transfer by matching deep CNN feature statistics, but their iterative optimization is slow. Subsequent work trained feed-forward networks for real-time stylization [20, 21], achieving similar quality via perceptual losses. Arbitrary style transfer meth-

ods then incorporated feature normalization tricks: AdaIN [22] aligns mean/variance in feature space, SaNet [23] and AdaAttN [24] perform feature statistics alignment on attention-weighted mean/variance and WCT [25] uses whitening/coloring transforms for universal stylization. Many extensions like multi-scale CNNs [26] have been proposed to improve content preservation and flexibility but fail to capture the nuances of identity representation in the stylized domain.

Parallel to CNN-based methods, GAN architectures [27] were applied to style transfer. Unpaired image-translation models like CycleGAN [28], MUNIT [29] and StarGAN [30] learn domain mappings without paired data, enabling photorealistic or artistic style transfers (e.g., photo-to-painting, day-to-night). More recently, large diffusion models have been used for style transfer. For instance, LSAST [31] uses step-aware prompts in a pretrained Stable Diffusion to impose fine-grained artistic styles while preserving content structure. Similarly, DiffStyler [32] uses a diffusion model with LoRA modules and mask-based injections to achieve localized, arbitrary style transfer. These diffusion-based methods can generate highly realistic stylizations, complementing earlier CNN/GAN approaches.

2.3 Cross-Domain Face Synthesis

Cross-domain face synthesis treats tasks like can be modelled as an image translation problems for tasks like photo-to-cartoon or photo-to-sketch. Early paired methods include pix2pix [33], which learns supervised translations, while CycleGAN [28] introduced cycle-consistency for unpaired domains. Extensions such as UNIT [34], MUNIT [29] and UGATIT [35] disentangle content and style to allow multimodal outputs. Even though training such generators is notoriously difficult, modern image to image translation methods [36, 37] have shown impressive results and tremendous potential.

In the specific context of faces, specialized networks have been developed: CartoonGAN [38] was one of the first GANs to convert photos into cartoon styles; CariGANs [39] targeted photo-to-caricature conversion by decoupling geometric exaggeration and appearance stylization. Other work focuses on face-to-sketch or face-to-anime translation using similar GAN architectures and carefully curated datasets. These models typically preserve facial structure (via losses or latent-sharing) while altering style. Recent diffusion-based approaches also support cross-domain translation. For example, pretrained text-to-image diffusion models can be guided (via conditioning or prompts) to translate a face into a different domain (e.g. sketch or cartoon) without task-specific retraining. Although not yet as common as GANs for cross-domain faces, diffusion pipelines (with control networks or specially trained prompts) demonstrate strong qualitative results for style-domain translations. Overall, cross-domain face synthesis has matured from supervised pix2pix networks [33] to versatile GANs [30, 34, 38] and even diffusion models for translating faces across diverse visual domains.

2.4 Face Swapping

Face swapping combines identity preservation with expression or pose transfer. Classical pipelines often relied on 3D face models, but modern approaches use deep networks. FSGAN [40] was among the first to propose subject-agnostic face swapping and reenactment: given a source and target face, FSGAN uses landmark-based reenactment and blending networks to swap identities without subject-specific training. Later works like FaceShifter [41] and SimSwap [42] improved photo-realism by disentangling identity and attributes via additional training losses. Diffusion models have also entered face swapping: DreamID [43] is a recent diffusion-based method that explicitly uses triplet-ID training for sharp, high-fidelity swaps in a single step. Across these methods, the key is enforcing the source identity while maintaining the target's pose/expression. In the realm of stylized faces, there is growing interest in combining style transfer with face reenactment. For example, ToonAging [44] jointly performs face re-aging and cartoon-style transfer, fusing age features with illustrative style in one generative step. While fully integrated stylized face reenactment remains an emerging area, initial works suggest that disentangled latent representations can blend identity, expression, and artistic style.

2.5 Efficient Generative Pipelines

Diffusion models achieve remarkable quality but are computationally intensive. A major research direction is accelerating inference and scaling up. Latent diffusion [45] is one strategy: it compresses images via a frozen VAE so that diffusion operates in a lower-dimensional latent space, drastically reducing compute compared to pixel-space diffusion. Other work optimizes the sampling process: for instance, Song and collaborators proposed DDIM sampling and related solvers to reduce the required denoising steps. DPM-Solver [46] achieves high-quality images (e.g. FID of about 4.7) with as few as 10–20 function evaluations, yielding upto 5× speedups. Salimans and Ho [47] introduce progressive distillation, training a small-step diffusion by distilling from a larger one (they report good results with only 4 steps on CIFAR-10). Architectural optimizations have also been pursued. DiffuSM [48] replace the standard attention-based U-Net with state-space-model layers, greatly reducing GFLOPs while matching image quality. Overall, efficient diffusion pipelines leverage latent representations, advanced ODE solvers, distillation, and novel model architectures to approach real-time performance without sacrificing fidelity.

These recent developments demonstrate that combining attention mechanisms, surrogate supervision, contrastive identity modeling, and diffusion generation significantly improves identity preservation in face swapping. Our work draws from these directions and applies them to the novel and challenging context of stylized scenic sketches, contributing both new architectural insights and a practical real-time implementation.

Chapter 3

Identity Distribution Analysis in StyleGAN-based Domain Transfer

In the pursuit of effective identity-preserving face swapping across domains, we first need to understand how identity representation varies between different visual domains. This chapter explores the relationship between facial identity representation and visual domains through a face cartoonization framework. Specifically, we investigate how facial identity information behaves when transformed from the photorealistic domain to the cartoon domain, with the hypothesis that identity features may manifest differently depending on the visual style.

3.1 Introduction

Cartoonization represents an important case study in cross-domain face synthesis, as it involves significant abstraction from photorealistic details while still needing to maintain recognizable identity. Unlike other artistic styles that add intricate textures like brush strokes or shading lines, cartoon images achieve their distinctive aesthetic through simplification and abstraction. Common features of cartoon images include well-defined edges, consistent smooth color shading, and relatively plain textures [38].

Our approach leverages the expressive latent space of StyleGAN [49] paired with a cartoon generator to transform facial images into the cartoon domain while preserving identity and pose information. By analyzing the behavior of identity embeddings before and after cartoonization, we demonstrate that identity distributions shift significantly when changing domains, revealing a fundamental correlation between identity representation and style domain.

3.2 Methodology

3.2.1 System Overview

Our cartoonization framework (Figure 3.1) takes two images as input: an identity image (I_{id}) and a pose image (I_p). The system extracts the required features from these images using dedicated encoders—E_{id} for identity and E_p for pose. These features are processed through a Multi-Layer Perceptron (MLP) that

generates a vector w in StyleGAN's latent space W, corresponding to the cartoonized version of I_{id} in the pose of I_p. This vector is then passed through a generator setup to produce the cartoon face output.

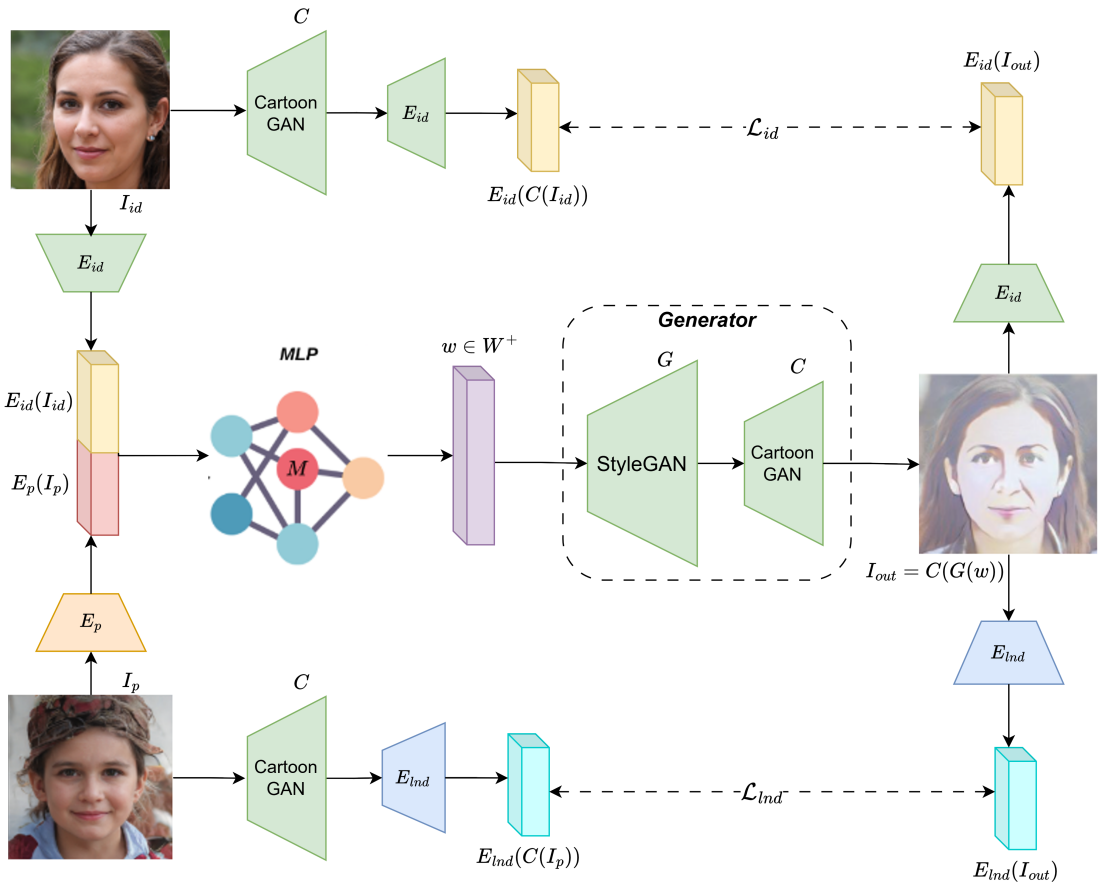


Figure 3.1: Our model architecture for face cartoonization. All models in green are pretrained, while the pose encoder in yellow is being fine-tuned and the MLP is trained from scratch. Data flow is marked with solid lines and losses are marked with dashed lines. Bold-dashed lines enclose the generator setup. Input images I_{id} and I_p are encoded using E_{id} and E_p respectively. The embeddings are concatenated and passed through the MLP which maps them to the W latent space of StyleGAN. The generated w vector is passed through our generator setup to give the final output $C(G(w))$.

3.2.2 Encoders

Our approach integrates identity and pose information using established models:

- **Identity Encoder** (E_{id}): We employ a pre-trained ArcFace [10] face recognition model, extracting activations from its penultimate layer to create an identity embedding vector.

- **Pose Encoder (E_p)**: For pose encoding, we use a pretrained VGG network [50] that has been trained on face images. This encoder is further fine-tuned during training to better capture pose information. The penultimate layer activations serve as our pose embedding vector.

The two embeddings are concatenated and passed through the MLP to generate the latent vector:

$$w = M([E_p(I_p), E_{id}(I_{id})]) \tag{3.1}$$

We also utilize a landmark encoder E_{lnd} during training to capture facial landmarks of realistic images, which helps ensure consistency in facial structure.

3.2.3 Generator Setup

Our generator comprises two essential components:

- **StyleGAN Generator (G)**: We employ a StyleGAN that has been pre-trained on the FFHQ dataset [49]. This functions as a pre-cartoonization backbone. Our aim is not to generate realistic faces but rather faces that will cartoonize well in the subsequent step.

- **Cartoon Generator (C)**: To transfer the style from realistic to cartoon, we pass our image through a GAN-based cartoon generator called White Box Cartoonizer [51]. This framework breaks down the cartoonization problem into surface representation, structure representation, and texture representation, optimizing for them individually to make the framework more controllable.

The latent vector w is passed through G and then through C to get our output image:

$$I_{out} = C(G(w)) \tag{3.2}$$

3.2.4 Loss Functions

To ensure proper disentanglement of identity and pose while maintaining cartoon aesthetics, we employ several specialized loss functions:

Identity Loss: To preserve the essential identity during cartoonization, we introduce an L_1 cycle consistency loss between the cartoonized identity image and our output:

$$L_{id} = \|E_{id}(C(I_{id})) - E_{id}(I_{out})\|_1 \tag{3.3}$$

This approach aligns with our objective of preserving identity attributes in the transformed cartoon space, even as the cartoonization process introduces stylistic exaggerations and distortions.

Landmark Loss: To ensure proper pose transfer, we incorporate a sparse L_2 cycle consistency landmark loss:

$$L_{lnd} = \|E_{lnd}(C(I_p)) - E_{lnd}(I_{out})\|_2 \tag{3.4}$$

This loss is crucial for accurate transfer of pose-related information, ensuring that the cartoonized output maintains the same facial arrangement as the pose reference image.

Reconstruction Loss: For preservation of other factors such as illumination and color, we use a mix loss L_{mix} [52] when the identity and pose images are identical:

$$L_{mix} = \alpha(1 - \text{MS-SSIM}(I_p, I_{out})) + (1 - \alpha)\|I_p - I_{out}\|_1 \tag{3.5}$$

where α is a hyperparameter. We impose a constraint on this loss, applying it exclusively when $I_{id} = I_p$:

$$L_{rec} = \begin{cases} L_{mix} & \text{if } I_{id} = I_p \\ 0 & \text{otherwise} \end{cases} \tag{3.6}$$

The overall loss is a weighted sum of the individual loss terms:

$$L_{total} = \lambda_1 L_{id} + \lambda_2 L_{lnd} + \lambda_3 L_{rec} \tag{3.7}$$

3.3 Experiments and Results

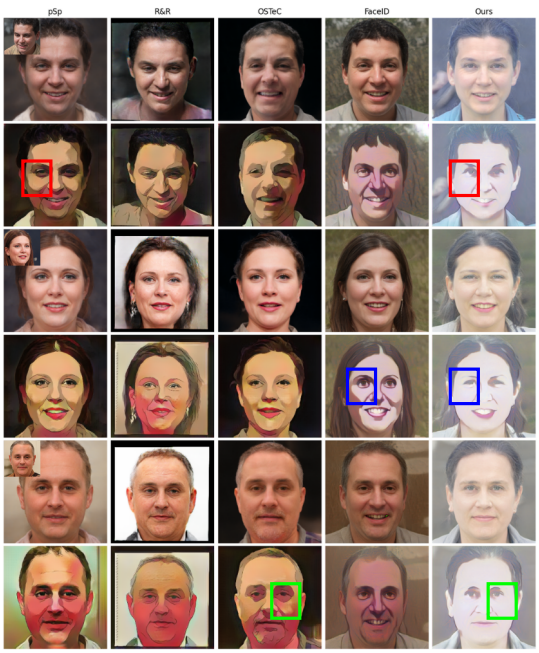
3.3.1 Implementation Details

For all experiments, we use StyleGAN pre-trained at 256×256 resolution. The ratio of training samples with $I_{id} = I_p$ and $I_{id} \neq I_p$ controls the weight for disentanglement and reconstruction. Following established practice, we use $I_{id} = I_p$ every third iteration, and $I_{id} \neq I_p$ otherwise. Loss weights are set to $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 0.001$ and $\alpha = 0.84$.

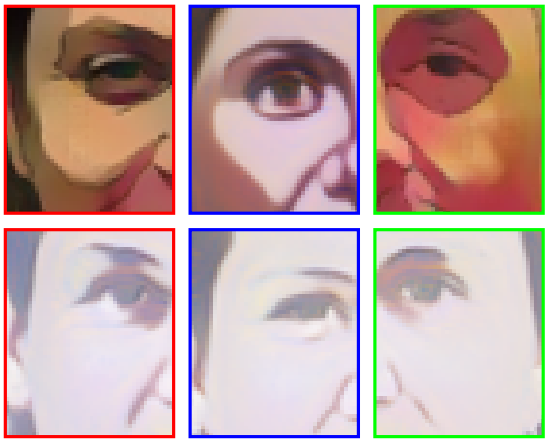
3.3.2 Domain-Specific Identity Distribution

To validate our hypothesis that identity representation is domain-dependent, we quantified the loss in identity information during the cartoonization process. We compared our results against those of other face frontalization models coupled with the White-Box Cartoonizer.

To calculate the loss in identity-related information after cartoonization, we subtracted the identity loss before cartoonization from the identity loss after cartoonization. Our findings, listed in Table 3.1 and seen in Figure 4.5, clearly show that our model exhibits superior performance in mitigating the identity loss increment during the cartoonization process. This result demonstrates the robustness of our approach and its effectiveness in maintaining identity-related information during the style transfer from realistic to cartoon aesthetics.



(a)



(b)

Figure 3.2: Comparison of our method with face frontalization models followed by successive cartoonization. Notice how our intermediary frontal images are better suited for cartoonization, producing results with well-defined edges, consistent smooth shading, and plain textures. Other methods often produce shading and texture artifacts, especially around the eyes, cheeks, and mouth regions.

3.3.3 Different W Spaces

To further validate that our encoder learns a novel distribution within the StyleGAN latent space that is better suited for cartoonization, we visually compared the original w vectors with those generated by our trained MLP using t-SNE visualizations [56].

As shown in Figure 3.3, the patterns and clusters formed by the StyleGAN-derived w vectors exhibit clear separation from those formed by the vectors generated by our MLP. This visual divergence strongly supports our hypothesis that the latent spaces of StyleGAN-generated images and the images produced by our encoder form distinctly different distributions for the same identities. This finding validates that facial identity representation is domain-dependent, which has significant implications for cross-domain identity preservation techniques.

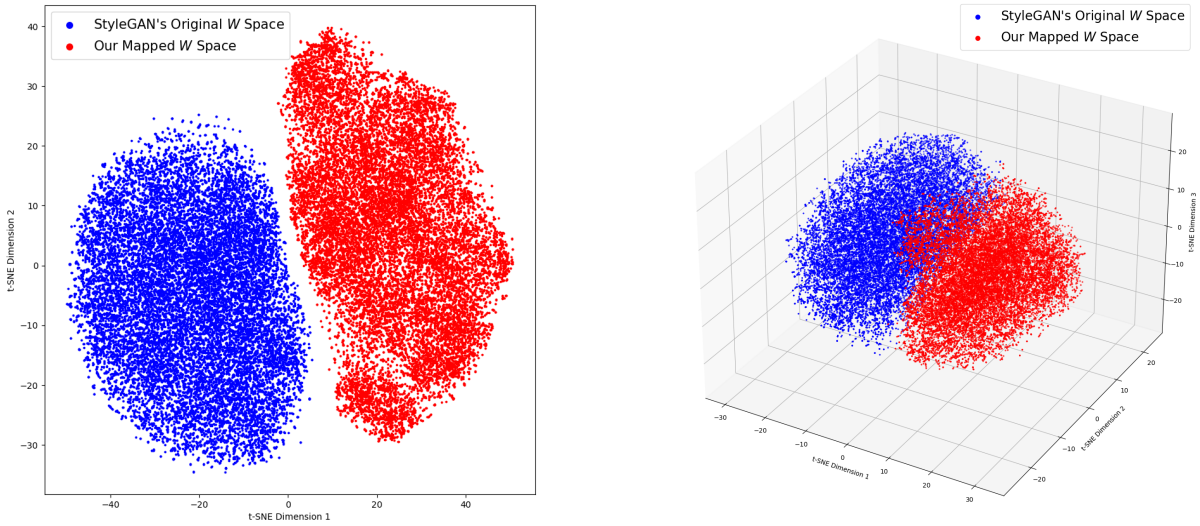
3.3.4 Ablation Studies

To demonstrate the effectiveness of each component in our method, we performed ablation experiments by removing individual loss components.

As observed in Figure 3.4, removing the reconstruction loss caused smaller details that contribute to identity to be lost, resulting in messier facial appearances. Without the landmark loss, the results

Method	**Average Identity Loss (L_{id}) ↓**		**Increment ↓**
	Before Cartoonization	After Cartoonization	
pSp [53]	0.1548	0.5092	0.3544
R&R [54]	0.3010	0.5797	0.2787
OSTeC [55]	0.2326	0.5710	0.3384
FaceID [52]	0.3441	0.7046	0.3605
Ours	0.4903	**0.6117**	**0.1213**

Table 3.1: Quantitative evaluation of frontalization and subsequent cartoonization on StyleGAN generated data with different identities and poses. Our method shows the lowest identity loss increment after cartoonization.



(a) 2-d t-SNE plot

(b) 3-d t-SNE plot

Figure 3.3: t-SNE visualizations of the latent vectors obtained from two distributions: StyleGAN's original W space and our mapped W space. The clear separation between clusters indicates that our model learns a distinct distribution that is better suited for identity preservation during cartoonization.

consistently exhibited the same pose regardless of the input pose image, indicating that this loss is crucial for proper pose transfer. We also experimented with adding an adversarial loss term using the randomly sampled w vectors from StyleGAN's latent space as real samples. However, this addition caused the outputs to retain substantial intricate details that contribute to heightened realism, compromising the desired smooth textures of cartoon aesthetics.

3.4 Conclusion

This chapter introduces a cartoonization technique integrated with a StyleGAN editing framework which uses identity and pose inputs to achieve face re-enactment in the cartoon domain. Identity preservation is enforced by a cross-domain identity loss causes the model to implicitly learn a different iden-

(a) results without $\mathcal{L}_{rec}$ (b) results with $\mathcal{L}_{adv}$ (c) results without $\mathcal{L}_{lnd}$

Figure 3.4: Results of ablation studies: (a) Without reconstruction loss, smaller details that contribute to identity are lost and the face appears messy. (b) Adding adversarial loss makes the output more realistic but compromises cartoon aesthetics. (c) Without landmark loss, outputs tend to have the same pose regardless of input.

tity transformation. We show that the encoder's learned distribution in W is more effective at capturing identity cues when the target domain is cartoonized. These results substantiate the hypothesis that facial identity representations are inherently domain-specific: the cues that define identity in photorealistic images are not identical to those in cartoon renditions.

The demonstrated domain dependence has significant repercussions for cross-domain face swapping. Rather than a naive transfer of identity embeddings, preserving identity across stylistic boundaries demands adaptation of the representation to domain-specific nuances. This insight acts as the foundation for our subsequent work on portrait sketch generation and efficient face swapping systems, explored in the following chapters.

Chapter 4

PS-StyleGAN: Illustrative Portrait Sketching using Attention-Based Style Adaptation

Portrait sketching represents a sophisticated artistic endeavor that requires capturing identity-specific attributes of a real face using abstract lines and shades. Unlike photorealistic image generation, effective portrait sketch synthesis necessitates selective attention to detail, making it particularly challenging for computational approaches. This chapter introduces Portrait Sketching StyleGAN (PS-StyleGAN), a novel style transfer approach specifically engineered for portrait sketch synthesis that builds upon our understanding of domain-dependent identity representation established in the previous chapter. The key innovation of PS-StyleGAN lies in its ability to leverage the semantic W^+ latent space of StyleGAN to generate high-quality portrait sketches while preserving the ability to make meaningful edits—such as pose and expression alterations—without compromising identity. This capability directly addresses our thesis's central concern of versatile identity preservation across visual domains. We achieve this through the implementation of Attentive Affine transform blocks in our architecture and a training strategy that utilizes transfer learning to modify StyleGAN's photorealistic outputs .

4.1 Background and Motivation

4.1.1 The Challenge of Portrait Sketch Generation

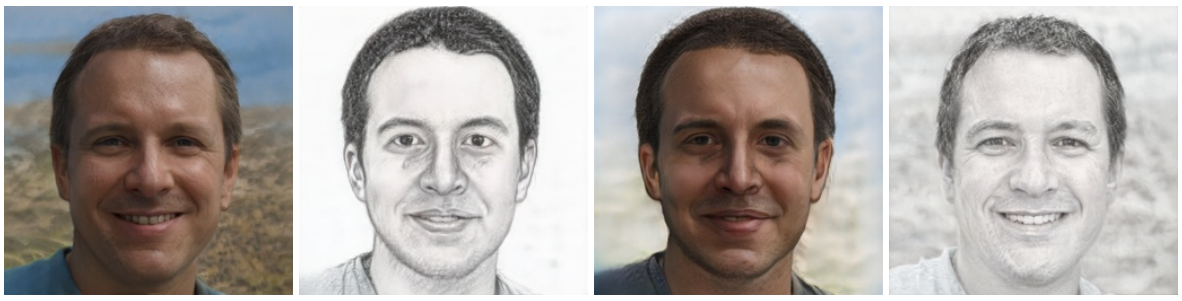
Portrait sketching presents unique computational challenges due to the abstract representation of facial features using minimal elements such as lines. Research into human perception of sketches [57, 58] has established that humans can recognize identities from simple line drawings, but the process by which artists select which lines to draw remains incompletely understood [59, 60]. Traditional non-photorealistic rendering methods [14–18] have attempted to address this artistic challenge but typically rely on ground truth geometry that is often noisy near detailed facial regions such as eyes, nose, and lips. This presents a significant limitation as humans are particularly sensitive to details in these regions due to specialized neural pathways for face detection and identification [61].

4.1.2 Deep Learning Approaches to Sketch Generation

The emergence of deep learning has catalyzed progress in sketch generation through style transfer [19] and image-to-image translation [28,33] approaches. Style transfer methods have shown success in global texture transformations but frequently struggle with preserving local details crucial for facial identity preservation. Conditional Generative Adversarial Networks (cGANs) [33, 62] have demonstrated promising results for general image-to-image translation, though training such generators remains challenging and typically requires large datasets—a constraint particularly limiting for facial sketch applications. More recent innovations have utilized pretrained StyleGAN models [49, 63] with encoders that invert images into StyleGAN's latent space to facilitate image-to-image translation [64]. The highly semantic W^+ latent space of StyleGAN enables meaningful edits to the output, such as changing pose or facial expression, without compromising identity. However, incorporating sketch styling into StyleGAN presents significant challenges, as it risks perturbing the latent space behavior of the pretrained model, potentially compromising semantic editing capabilities.

4.1.3 Limitations of Current Approaches

Previous methods like DualStyleGAN [65] have attempted to address this challenge by disentangling StyleGAN's spatial resolution layers to perform independent structure and color transfer between domains. Their approach incorporates ResBlock-based feature statistics alignment using AdaIN [22] that provides structure control over coarse and middle layers of StyleGAN. While this approach preserves the latent distribution to enable semantic editing, DualStyleGAN's style blending can result in significant identity loss as it does not account for domain-specific identity nuances. Our experimental analysis (shown in Figure 4.1) reveals that structure and color entanglement exists across all layers of StyleGAN, particularly in the middle layers. This entanglement complicates the decoupling of structure and color transformations without sacrificing artistic characteristics such as pencil strokes and shading.



(a) input (b) sketch (c) structure only (d) color only

Figure 4.1: Results of DualStyleGAN trained on CUHK dataset. The generated sketch (b) results from complete structure and color transfer. Structure transfer (c) shows considerable identity loss, while color transfer (d) fails to achieve proper stylization.

4.2 Methodology

4.2.1 Overview of PS-StyleGAN

To address the limitations of existing approaches, we propose Portrait Sketching StyleGAN (PS-StyleGAN), which converts real face photos into portrait sketches while maintaining StyleGAN's semantic editability without requiring generator fine-tuning. Our approach employs novel attention-based affine transformation blocks to modulate only the style latent codes while keeping the generator frozen. These blocks effectively simulate the behavior of a fine-tuned StyleGAN. By eschewing structure transfer entirely, we ensure identity preservation and adopt a progressive training strategy to achieve rapid yet smooth domain transfer. Our model demonstrates inversion consistency across different datasets, allowing for reliable manipulation of the generated sketches.

4.2.2 System Architecture

PS-StyleGAN operates as an end-to-end system for facial sketch synthesis, as illustrated in Figure 4.2. Given a content image C and a sketch image S representing a particular style $\mathcal{S}$, we first invert both images to the Z^+ latent space of a pre-trained StyleGAN generator g using a pSp-based encoder E [53,66]. The encoder is trained on 256×256 resolution images from the FFHQ dataset [49] and modified to embed face images into the Z^+ latent space, which demonstrates greater resilience to background details compared to the standard W^+ space. Using StyleGAN's mapping network f, we transform these embeddings into latent codes w_c^+ and w_s^+ within the shared W^+ latent space of g. Finally, we pass these latent codes through our novel synthesis network g' to generate the output image G, which successfully captures the style of S while preserving the identity of C.

4.2.3 Hierarchical Style Control in StyleGAN

StyleGAN's architecture provides natural hierarchical control over different aspects of image generation through its style blocks at varying spatial resolutions:

- **Coarse layers** ($4 \times 4 - 8 \times 8$ resolution) control high-level aspects such as pose, face structure, hair texture, and accessories
- **Middle layers** ($16 \times 16 - 32 \times 32$ resolution) generate smaller-scale features like eyes, smile, and hairstyle
- **Fine layers** ($64\times64-256\times256$ resolution) primarily control the color scheme and microstructure

To address the challenges identified earlier, we implement attention-based style adaptation blocks specifically in the fine layers of the generator network. These blocks perform feature transformations that consider both global and local style patterns. Each block consists of a novel Attentive Affine transform module and StyleGAN's modulative convolution layer, providing instance-wise style conditioning to the

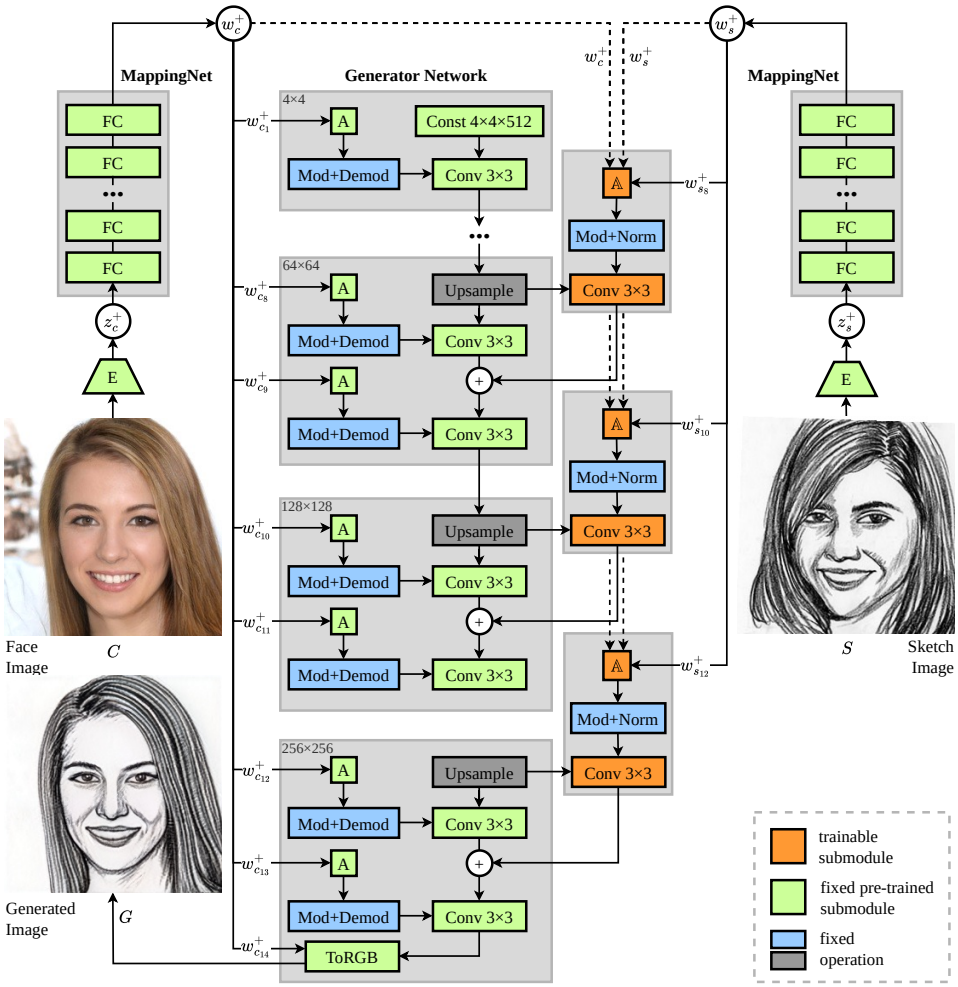


Figure 4.2: Overview of the PS-StyleGAN architecture. The model uses a pretrained 256x256 resolution StyleGAN2 generator fitted with three style adaptation blocks at the fine resolution layers. Each block contains a novel Attentive Affine transform module that predicts affine parameters from attention-weighted latent codes to modulate spatial features at different scales.

content features. By restricting modulation to fine layer features, we preserve the overall structure of the content image. The adapted features are then fused with the original content features at each layer, facilitating a smooth transition from the photorealistic domain to the sketch domain. Our experimental results confirm that this approach maintains the consistency of StyleGAN's latent space, enabling continued manipulation of sketches using methods designed for realistic images.

4.2.4 Attention-Based Latent Space Transformation

4.2.4.1 Attentive Affine Transformations

Inspired by AdaAttN [24], we introduce attentive affine transformations to obtain improved affine parameters y_i^S at the fine layers, which encapsulate the complete feature distribution of the style image.

These parameters are subsequently used by the Adaptive Instance Normalization (AdaIN) operation to achieve style transfer. The style adaptation process operates in three steps:

1. Computing attention maps using content and style latent codes w^+c and w^+s
2. Calculating a weighted segment of the style latent code and deriving improved affine parameters $y^S s, i$ and $y^S b, i$ for the style features
3. Adaptively normalizing the content features for instance-wise feature distribution alignment

4.2.4.2 Attention Map Generation

Unlike standard style transfer methods that measure similarity between content and style features, we employ the attention mechanism to assess similarity between content and style latent codes. This approach exploits the highly disentangled nature of StyleGAN's W^+ latent space, where similarity in latent space effectively extends to feature space similarity. Additionally, the relatively low dimensionality of the latent space keeps the model lightweight and reduces the computational cost of attention map calculation. To compute the attention map A corresponding to fine layer i, we formulate query (Q), key (K), and value (V) as:

$$Q = f(\text{Norm}(w_c^+))\ K \qquad = g(\text{Norm}(w^+ s))\ V = h(w^+ s, i) \tag{4.1}$$

where f, g, and h are standard trainable 1×1 convolution layers, and Norm represents instance normalization applied channel-wise. We then compute the attention map A as:

$$A = \text{Softmax}(Q^T \otimes K) \tag{4.2}$$

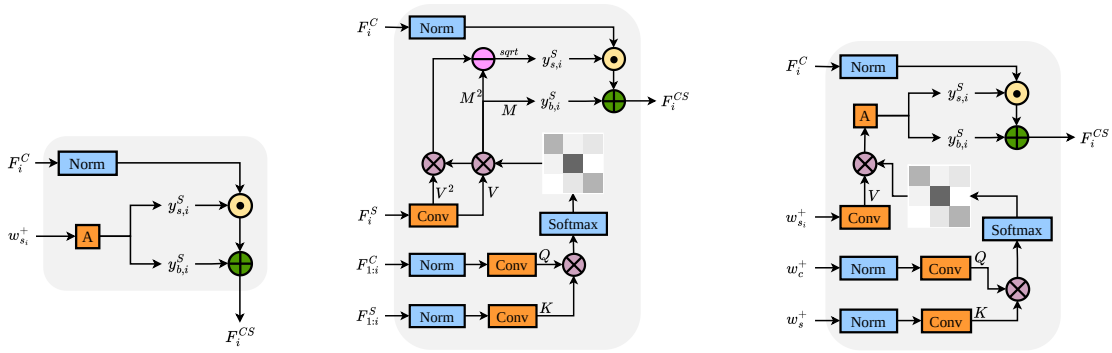
where $\otimes$ represents matrix multiplication.

4.2.4.3 Improved Affine Parameters

While AdaAttN applies the attention map to style features, our approach applies it to style latent code segments. The style latent code segment $w^+ s, i$ is multiplied with the attention score matrix to represent it as a distribution of all style points in the latent space. We refer to this as the attention-weighted latent code segment $x^+ s, i \in \mathbb{R}^{512}$, from which we learn improved affine transformations to derive more representative affine parameters $y^S_{s,i}$ and $y^S_{b,i}$:

$$x^+_{s,i} = V \otimes A^T\ (y^S_{s,i}, y^S_{b,i}) \qquad = \text{Affine}(x^+_{s,i}) \tag{4.3}$$

where Affine is a learnable single fully-connected layer identical to StyleGAN's traditional affine transform. The output dimensionality of this layer is twice the number of feature maps at the corresponding spatial resolution of the generator.



(a) AdaIN (b) AdaAttN (c) Ours

Figure 4.3: (a) The structure of AdaIN [22] module used in StyleGAN [49]. (b) The structure of AdaAttN [24] module. (c) The structure of our proposed design showing *attentive affine* transform blocks. Here, *A* denotes a basic affine transform block consisting of a single trainable fully-connected layer and *Norm* denotes channel-wise mean-variance normalization.

4.2.4.4 Adaptive Normalization

Finally, we use the derived affine parameters to modulate the normalized content feature map pointwise for each channel, generating the transformed feature map. The AdaIN operation in our implementation becomes:

$$F_i^{CS} = y_{s,i}^S \frac{F_i^C - \mu(F_i^C)}{\sigma(F_i^C)} + y_{b,i}^S \tag{4.4}$$

The transformed feature maps F_i^{CS} are then processed through a trainable convolution layer, whose outputs are selectively fused with those of the fine layers from the pre-trained synthesis network g to complete the style adaptation process. Notably, we observed that omitting the mean affine parameter $y_{b,i}^S$ during modulation does not significantly affect the generated results. Therefore, following StyleGAN2's approach, we combine the modulation and convolution operations by scaling the convolution weights, effectively reducing the output dimensionality of the affine transform blocks. In summary, our method performs feature statistics alignment using attentive affine transformations by generating attention-weighted latent codes that better represent the target style feature distribution in the fine layers, ensuring preservation of middle and coarse layer features crucial for identity preservation. This ensures that the identity transformations unique to the style are learned.

4.2.5 Training Strategy

We employ a progressive transfer learning scheme using a pretrained StyleGAN to gradually refine its generative space to align with the target style distribution $\mathcal{S}$, which may comprise limited samples. This scheme consists of two stages, as illustrated in Figure 4.4.

4.2.5.1 Stage I: Domain Transfer

Similar to fine-tuning, we aim to achieve a general transformation from the photorealistic domain to the sketch domain defined by $\mathcal{S}$. We randomly generate a latent code z^+ and sample a sketch image S with its corresponding style latent code z_s^+. Using StyleGAN's mapping network f, we obtain the W^+ latent space embeddings for the content and style images as $w^+ = f(z^+)$ and $w_s^+ = f(z_s^+)$, respectively. We then pass these latent codes through our synthesis network g' to obtain the generated sketch $G = g'(w^+, w_s^+)$. Following standard style transfer practices, we employ a style loss to fit the style of the generated sketch G to S:

$$\mathcal{L}_{sty} = \lambda_{CX}\mathcal{L}_{CX}(G, S) + \lambda_{FM}\mathcal{L}_{FM}(G, S) \tag{4.5}$$

where $\mathcal{L}_{CX}$ denotes contextual loss [67] and $\mathcal{L}_{FM}$ denotes feature matching loss [22]. To preserve content features, we use an identity loss [10] between G and the reconstructed content image $g(w^+)$, constituting a content loss:

$$\mathcal{L}_{cont} = \lambda_{ID}\mathcal{L}_{ID}(G, g(w^+)) \tag{4.6}$$

Adding the standard StyleGAN adversarial loss $\mathcal{L}_{adv}$, our complete objective function for Stage I becomes:

$$\min_G \max_D \lambda_{adv}\mathcal{L}adv + \mathcal{L}sty + \mathcal{L}_{cont} \tag{4.7}$$

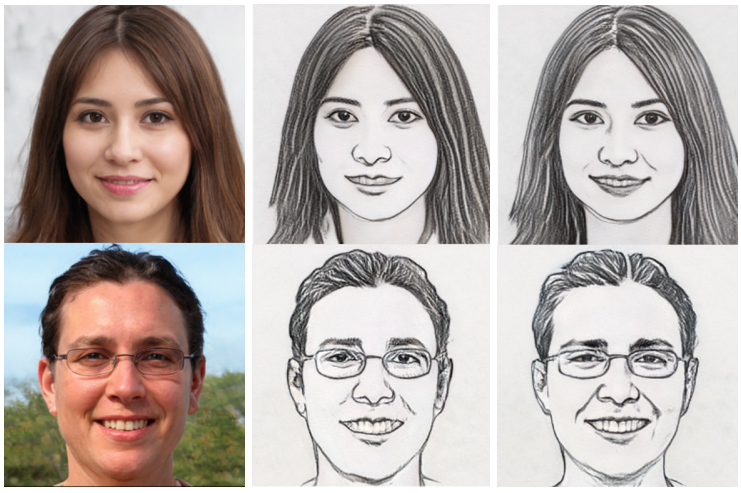
4.2.5.2 Stage II: Conditional Refinement

Stage I transforms StyleGAN's generative space to a narrow domain, which limits its ability to capture the diverse styles contained in $\mathcal{S}$, as shown in Figure 4.4(b). To address this limitation, we use paired data of ground truth sketches and their photorealistic counterparts as conditional supervision to broaden the generative domain. Given a sketch image S and corresponding photo P, we obtain the W^+ latent space embeddings as $w_s^+ = f(E(S))$ and $w_p^+ = f(E(P))$, and use them to generate the sketch $G = g'(w_p^+, w^+s)$. In addition to the losses used in Stage I, we employ perceptual loss [20] for G to reconstruct S, thereby learning a varied set of style-specific transformations. We also introduce a regularization term in $\mathcal{L}_{cont}$, which is the L_2 norm of the convolution weights comprising our style adaptation blocks:

$$\mathcal{L}_{cont} = \lambda_{ID}\mathcal{L}_{ID}(G, g(w^+p)) + \lambda reg||W||^2 \tag{4.8}$$

where W represents the weight matrices of the trainable convolution layers. This regularization term controls the degree of style adaptation and helps prevent overfitting. Thus, the objective function for Stage II becomes:

$$\min_G \max_D \lambda_{adv}\mathcal{L}_{adv} + \lambda perc\mathcal{L}_{perc} + \mathcal{L}_{sty} + \mathcal{L}_{cont} \tag{4.9}$$



(a) input (b) after stage I (c) after stage II

Figure 4.4: Results after each stage of progressive transfer learning. Stage I converges to an average representative style (b) where facial features are sketched similarly across samples. Stage II widens the model's generative space to capture subtle style variations resulting in better identity preservation (c).

4.3 Experimental Results

4.3.1 Datasets

We conducted comprehensive evaluations on multiple datasets to assess our method's effectiveness:

- **FS2K** [68]: The most extensive publicly available Face Sketch Synthesis (FSS) dataset, comprising 2,104 photo-sketch pairs with diverse image backgrounds, skin tones, sketch styles, and lighting conditions. The sketches are classified into three distinct artistic styles.
- **CUHK** [69]: A dataset predominantly featuring Asian faces, used to evaluate our method against DualStyleGAN, which introduces shape characteristic biases in its results.
- **APDrawing** [70]: A challenging dataset used to evaluate our method's ability to generalize and adapt to complex sketching scenarios.

4.3.2 Quantitative Evaluation

To quantitatively compare PS-StyleGAN with other state-of-the-art methods, we utilized four performance metrics:

- **Learned Perceptual Image Patch Similarity (LPIPS)** [71]: Measures perceptual similarity between generated sketches and ground truth; lower values indicate more realistic synthesis
- **Structure Co-Occurrence Texture (SCOOT)** [72]: Evaluates sketch quality based on structural and textural similarities; higher values indicate better resemblance to artist-drawn sketches

- **Feature Similarity Measure (FSIM)** [73]: Assesses low-level feature similarity; higher values indicate better correspondence with ground truth
- **ID loss** [10]: Quantifies identity preservation; lower values indicate better identity maintenance

Table 4.1 presents the average metric values across all test samples, demonstrating PS-StyleGAN's superior performance compared to current state-of-the-art methods.

Table 4.1: Quantitative comparison of PS-StyleGAN with HIDA [74], FSGAN [68], AdaAttN [24] and DualStyleGAN [65] based on SCOOT, LPIPS, FSIM and ID loss metrics.

Method	SCOOT ↑	LPIPS ↓	FSIM ↑	ID ↓
HIDA	0.4433	0.3214	0.3660	0.0241
FSGAN	0.3621	0.2890	0.3692	0.0424
AdaAttN	0.4670	0.2600	0.3806	0.0233
DualStyleGAN	0.4490	0.3012	0.3631	0.0247
Ours	**0.5603**	**0.2303**	**0.4283**	**0.0206**

4.3.3 Qualitative Analysis

Visual comparison of PS-StyleGAN with leading methods—FSGAN [68], HIDA [74], DualStyleGAN [65], and AdaAttN [24]—reveals that our method excels in rendering facial features, particularly eyes and lips, with sharper details and enhanced realism, as shown in Figure 4.5. While our results appear visually similar to DualStyleGAN, their method often learns shape biases present in the training dataset, which can affect identity recognition. DualStyleGAN frequently alters gaze direction and lip shape, compromising identity preservation. The Attentive Affine transform blocks in PS-StyleGAN contribute to a superior balance between artistic expression and accuracy, resulting in more visually appealing and faithful representations of facial features.

4.4 Applications to Identity-Preserving Face Swapping

PS-StyleGAN's ability to generate high-quality portrait sketches while preserving identity makes it particularly valuable for cross-domain face swapping applications. The model's capacity to maintain semantic editability of the latent space enables meaningful manipulations of pose, expression, and style while preserving core identity features. For the purpose of identity-preserving face swapping in scenic sketches, PS-StyleGAN offers several key advantages:

1. **Identity Consistency**: The model maintains facial identity across domain transformations, a prerequisite for convincing face swapping
2. **Style Flexibility**: With minimal paired examples (100), PS-StyleGAN can adapt to various artistic sketch styles

Figure 4.5: Comparison of our method with other state-of-the-art methods on the three styles of FS2K: style 1 (row 1), style 2 (row 2), style 3 (row 3). From left to right: Input identity image, Ours, DualStyleGAN, HIDA, FSGAN, AdaAttN.

3. **Efficient Training**: The rapid training time and lightweight architecture facilitate practical deployment

4. **Semantic Control**: The preserved latent space editability allows fine-grained control over facial attributes in the target domain

These capabilities lay the foundation for our subsequent chapter, which will explore diffusion-based methods for end-to-end face swapping across domains.

4.5 Limitations and Future Work

Despite its effectiveness, PS-StyleGAN exhibits certain limitations that present opportunities for future research:

- **Data Bias Susceptibility**: The model may inherit biases present in training datasets, potentially affecting generalization across diverse demographics
- **Style Variation Constraints**: While improvements were made through conditional refinement, extreme style variations remain challenging
- **Accessory Generation**: The current implementation struggles to generate realistic accessories in synthesized sketches

Future work could address these limitations by incorporating more diverse training data, exploring multi-modal style representation, and enhancing the model's capability to handle accessories and non-facial elements.

4.6 Conclusion

This chapter introduced Portrait Sketching StyleGAN (PS-StyleGAN), an approach specifically designed for the intricate color transformation requirements of portrait sketch synthesis. By leveraging StyleGAN's semantic W^+ latent space, our method not only generates high-quality portrait sketches but also preserves the ability to make meaningful edits while maintaining identity—a critical capability for face swapping applications. The incorporation of Attentive Affine transform blocks enables adaptation of StyleGAN outputs in an inversion-consistent manner by considering both content and style latent features. The model demonstrates remarkable effectiveness with minimal paired examples (around 100) and features a short training time, enhancing its practical applicability. PS-StyleGAN represents a significant advancement in portrait sketch generation that directly contributes to our thesis's goal of efficient identity-preserving face swapping across domains. The ability to maintain identity while transforming visual representation provides a crucial foundation for the diffusion-based approaches we will explore in the next chapter.

Chapter 5

High Fidelity Portrait Sketching using Diffusion

The core characteristic of high fidelity portrait sketch generation is identity preservation. As mentioned in chapter 4, GAN-based methods inherently suffer from the inversion problem that translates to a considerable loss of identity when it comes to faces that reside outside the domain of training data. Furthermore, the generative capabilities of StyleGAN is outmatched by diffusion models in terms of quality and control. This motivates our exploration of diffusion models. We adopt **InstantID** [75], a zero-shot identity-preserving framework, and extend it through multi-stage personalization to bridge the photo-to-sketch domain gap while preserving facial identity.

5.1 Overview of InstantID

5.1.1 Introduction

Methods for identity transfer in personalized image generation generally fall into two main approaches when using reference images. The first requires test-time fine-tuning (e.g., DreamBooth [76], Textual Inversion [77]) and achieves high fidelity but is computationally expensive, time-consuming, and demands multiple reference images—making it unsuitable for limited-data scenarios. The second category avoids fine-tuning by using lightweight adapters (e.g., IP-Adapter [78]) to inject visual features via cross-attention. However, these often rely on CLIP [12] encoders, which provide weakly aligned signals and struggle with identity fidelity.

InstantID bridges this gap by offering a plug-and-play solution that requires only a single reference image. It combines strong semantic features and minimal spatial constraints through its novel IdentityNet architecture, effectively guiding generation with identity embeddings, coarse facial landmarks, and optional text prompts.

5.1.2 Background Concepts

Diffusion. Stable Diffusion [79] performs efficient diffusion in low-dimensional latent space rather than pixel space using an auto-encoder architecture. For an input image $x_i \in R^{H\times W\times 3}$, the encoder maps it to a latent representation: $z_0 = \xi(x_i)$, where $z_0 \in R^{h\times w\times c}$ where the downsampling factor is

$f = H/h = W/w$ and c represents the latent dimension. A denoising U-Net [80] (ϵ_θ) is used to denoise normally-distributed noise ϵ with noisy latent z_t, current timestep t, condition C (text embeddings from CLIP).

The training objective minimizes the difference between the noise and the model's prediction:

$$\mathcal{L} = \mathbb{E}_{z_t,t,C,\epsilon\sim\mathcal{N}(0,1)}[||\epsilon - \epsilon_\theta(z_t, t, C)||_2^2] \tag{5.1}$$

Controllable Diffusion Models. ControlNet [81] is used for adding spatial control to pre-trained diffusion models, extending beyond textual prompts capabilities. It integrates the U-Net architecture from Stable Diffusion with a trainable replica that includes zero convolution layers in both encoder and middle blocks. It encodes spatial condition information (such as sketches, poses, or depth maps) by adding residuals to U-Net blocks and embedding this information into the original network.

$$y_c = \mathcal{F}(x, \theta) + \mathcal{Z}(\mathcal{F}(x + \mathcal{Z}(c, \theta_{z1}), \theta_c), \theta_{z2}) \tag{5.2}$$

Where $\mathcal{F}$ is the U-Net architecture, x is the latent, θ represents frozen weights from the pre-trained model, $\mathcal{Z}$ corresponds to zero convolutions with weights, θ_{z1}, θ_{z2}, and θ_c is the trainable weight of ControlNet.

Image Prompt Adapter IP-Adapter [78] enables image prompt capabilities alongside text prompts without modifying the original text-to-image models. It uses a decoupled cross-attention approach, embedding image features through additional cross-attention layers while preserving other parameters:

$$Z_{new} = Attention(Q, K^t, V^t) + \lambda \cdot Attention(Q, K^i, V^i) \tag{5.3}$$

Where Q, K^t, V^t are query, key, and value matrices for text cross-attention, K^i, V^i are for image cross-attention, λ controls the influence of the image prompt

5.1.3 Architecture

InstantID achieves its efficient, high identity preservation through the below three components.

Identity Encoder. Prior methods used CLIP encoder for visual prompt extraction. CLIP's key limitation stems from its weakly aligned training data, causing its encoded features to primarily capture vague semantic information (e.g., composition, style, colors). While useful as general supplements to text embeddings, these features lack the strong semantics and high fidelity needed for precise ID preservation tasks. It instead leverages pre-trained face recognition models to extract strong identity embeddings.

Image Adapter. To incorporate image prompts, IP-Adapter's lightweight design with decoupled cross-attention is used. Unlike IP-Adapter, InstantID instead utilizes the face ID embeddings discussed previously to derive semantic facial characteristics. These features are then transformed into the text feature space through a trainable projection layer, with the resulting output serving as the face embedding.

IdentityNet. Enhancing the image features may lead to impairment of the editing ability of the text tokens despite the decoupled cross-attention used. Thus, an additional set of image features are introduced through a modified ControlNet layer. Firstly, the detailed facial landmarks are replaced with five coarse keypoints (eyes, nose, mouth) to provide light spatial guidance without over-constraining pose or expression; and secondly, the text conditioning is replaced with identity embeddings in the cross-attention layers to focus generation solely on facial identity.

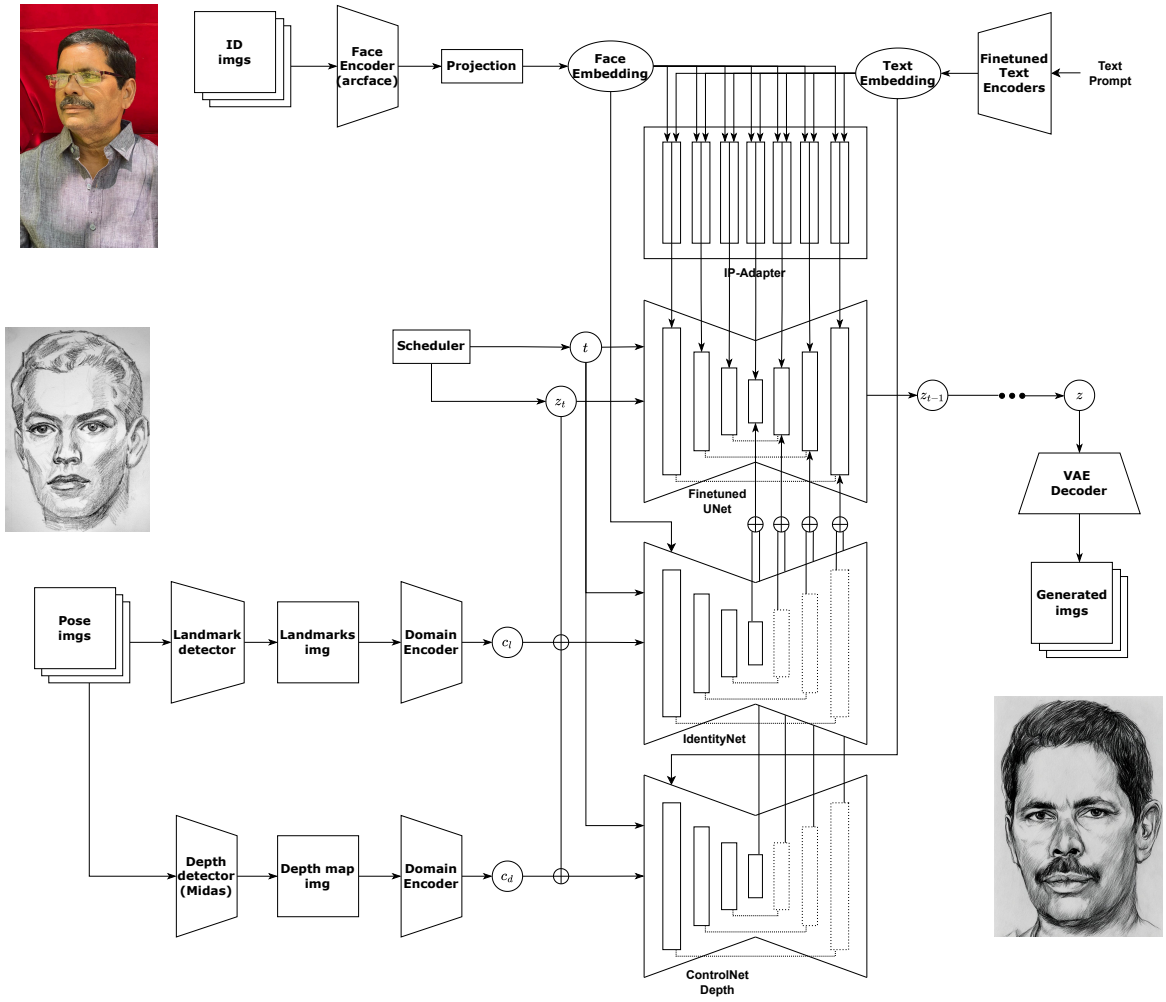


Figure 5.1: Overview of the InstantID architecture

5.1.4 Modifications

Our approach enhances InstantID by integrating an additional ControlNet layer for depth information processing using the MiDaS detector [82] as a preprocessor. While InstantID's IdentityNet effectively preserves core identity features, the sparse nature of identity reference points fails to capture the complete three-dimensional structure of facial features such as expressions. Our depth-enhanced

architecture works in parallel with IdentityNet to provide additional spatial guidance. The depth map ControlNet captures subtle facial contours missed by identity embedding and IdentityNet, resulting in more accurate preservation of structural characteristics. This integration significantly improves expression rendering and enables superior modelling of accessories like jewellery, glasses, and clothing details that are typically poorly represented. By combining IdentityNet's weak spatial guidance with additional depth-based spatial guidance, our model maintains geometric consistency across different poses and viewing angles, outperforming the standard InstantID pipeline in scenarios requiring high fidelity to identity and structural details.

5.2 Finetuning for Personalization

5.2.1 Background

While InstantID demonstrates strong identity preservation capabilities, its performance degrades when applied to niche domains (that don't exist in it's knowledge base) like sketches - a limitation stemming from subtle but critical differences in how identity is represented across visual domains as shown in the earlier sections.

To address this, we incorporate methods like DreamBooth [76] and Textual Inversion [77], which implicitly learn domain-specific identity transformations during their adaptation process. Following this principle, we implement targeted fine-tuning of InstantID's U-Net component using just 40 sketch samples. This lightweight adaptation allows the model to learn the unique characteristics of sketch-based identity representation while maintaining its core preservation capabilities.

Our approach consists of two seperate fine-tuning methodologies- DreamBooth and Textual Inversion-made efficient through LoRA. This multi-strategy adaptation framework allows the model to effectively bridge the representation gap between photorealistic domains and specialized artistic styles without compromising the core identity preservation mechanisms of InstantID.

Low-Rank Adaptation (LoRA). LoRA [83] enables parameter-efficient fine-tuning of large-scale models by decomposing weight updates into low-rank matrices. Instead of updating full attention weights $W \in \mathbb{R}^{d\times d}$, LoRA introduces trainable matrices $A \in \mathbb{R}^{d\times r}$ and $B \in \mathbb{R}^{r\times d}$, where $r \ll d$, and injects them as a residual:

$$W' = W + \alpha BA \tag{5.4}$$

Here, W remains frozen while only A and B are trained, with α scaling the update magnitude. This allows the model to specialize on new domains with significantly fewer trainable parameters and reduced overfitting.

Textual Inversion. Textual Inversion [77] introduces new pseudo-words into the text encoder vocabulary to represent novel visual concepts. A new token τ is initialized and optimized by minimizing the denoising loss while keeping all other model weights frozen. The learned token embedding e_τ captures the semantic and stylistic characteristics of the concept when used in prompts:

$$\min_{e_\tau} \mathbb{E}_{z_t,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, e_\tau)\|_2^2\right] \tag{5.5}$$

This allows concept-specific conditioning in the latent space without altering the diffusion model's internal parameters.

DreamBooth. DreamBooth [76] fine-tunes diffusion models for subject-driven generation by associating a unique identifier token with a target visual concept using class-specific prompts (e.g., 'a photo of [V] dog"). The model is fine-tuned on a few subject-specific images while regularized with prior-preserving loss to retain generalization:

$$\mathcal{L} * \text{total} = \mathcal{L} * \text{recon}(x\_\text{subject}, [V]c) + \lambda\mathcal{L} * \text{prior}(x * \text{class}, c) \tag{5.6}$$

where c is the class label and $[V]$ is the unique token. This allows the model to learn instance-level identity preservation while maintaining class-level semantics.

5.2.2 Method

DreamBooth + LoRA. To adapt the InstantID backbone to a niche sketch domain, we finetune the cross-attention layers of the U-Net using DreamBooth paired with Low-Rank Adaptation (LoRA). DreamBooth enables subject-specific tuning by conditioning the model on a novel concept token, ensuring identity consistency. LoRA enables efficient adaptation by injecting low-rank trainable matrices into attention projections while keeping the original model weights frozen. This combination allows the model to internalize sketch-specific structure and texture variations with minimal parameter updates, preserving general generation capabilities.

Textual Inversion + LoRA (Pivotal Tuning). To reinforce concept grounding in the text encoder, we use pivotal tuning [84] — an approach that first learns new token embeddings via Textual Inversion before LoRA fine-tuning. In this two-stage setup, token embeddings are optimized to represent the novel subject in the text space, after which U-Net adaptation via LoRA is performed while keeping embeddings fixed. This separation ensures that semantic alignment is learned prior to visual adaptation, improving prompt controllability and mitigating overfitting on small datasets.

While DreamBooth and Textual Inversion typically suffer from identity drift and require careful balancing to avoid overfitting, this use of InstantID as a backbone of this setting mitigates these issues through its dual-encoder architecture that anchors generation with strong identity embeddings. Because our fine-tuning targets only the U-Net component and is applied on top of InstantID's identity-locked backbone, these parameter-efficient methods serve as effective adapters rather than full retrainers, avoiding the catastrophic forgetting and semantic entanglement typically associated with such techniques.

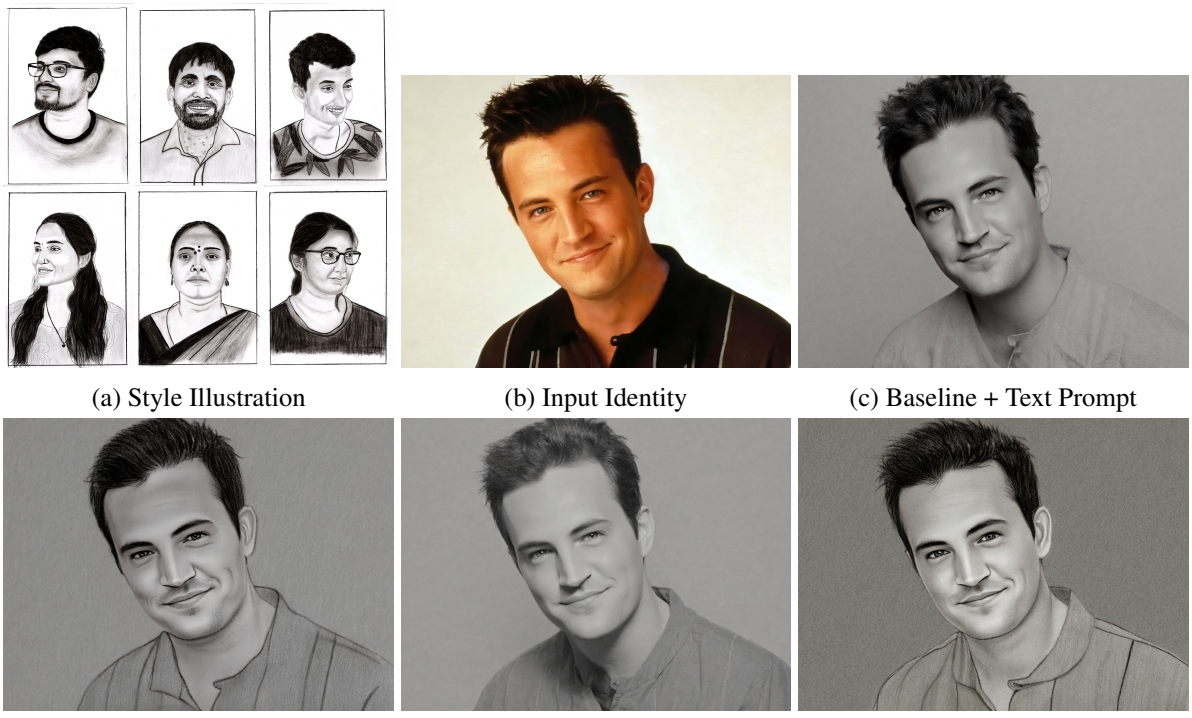
5.3 Results

We evaluate the performance of our approach through qualitative and quantitative comparisons with several baseline methods. The following sections present the results of our experiments.

5.3.1 Qualitative Comparison

We compare the generated images from the following fine-tuning approaches in Figure 5.2:

1. Baseline + Text Prompt
2. Baseline + Dreambooth-LoRA
3. Baseline + Pivotal Tuning
4. Baseline + Dreambooth-LoRA + Pivotal Tuning



(a) Style Illustration (b) Input Identity (c) Baseline + Text Prompt

(d) Baseline + Dreambooth-LoRA (e) Baseline + Pivotal Tuning (f) Baseline + Dreambooth-LoRA + Pivotal Tuning

Figure 5.2: Qualitative comparison of fine-tuning methods. Our combined approach (d) demonstrates superior performance in preserving identity and adapting style.

5.3.2 Quantitative Comparison

We report two key metrics—Fréchet Inception Distance (FID) and Identity (ID) Loss—to assess image quality and identity preservation.

Table 5.1: Quantitative comparison of different methods. Lower values indicate better performance for both metrics.

Method	FID ↓	ID Loss ↓
Baseline + Text Prompt	121	0.81
Baseline + Dreambooth-LoRA	93	0.78
Baseline + Pivotal Tuning	86	0.79
Baseline + Dreambooth-LoRA + Pivotal Tuning	**75**	**0.78**

The results indicate that our combined approach of Baseline + Pivotal Training + Dreambooth-LoRA achieves the best performance in terms of FID, confirming its effectiveness in generating high-quality, style-adherent images. The reported ID-loss values indicate that fine-tuning methods don't affect the identity-preserving nature of InstantID.

5.3.3 Conclusion

In this chapter, we explored how diffusion-based models can be adapted for generating high-quality, identity-preserving portrait sketches. By extending the InstantID framework to work in the sketch domain, we showed how strong identity cues can guide the generative process without the need for fine-tuning on individual users. This approach allows for zero-shot personalization, making it flexible and scalable for real-world use. Further, we show that custom style adaptation can be achieved using a combination of fine-tuning methods without compromising facial identity fidelity. Therefore, unlike traditional methods that struggle with maintaining identity in abstract representations, our modified diffusion model strikes a balance between artistic style and facial fidelity. These insights set the stage for the next chapter, where we move beyond individual components and bring everything together into a real-time, end-to-end face swapping system designed for tourist applications.

Chapter 6

Efficient Cross-Domain Face Swapping System for Tourist Applications

The tourism industry has continuously evolved to incorporate technological advancements that enhance visitor experiences. This chapter presents a novel application of face swapping technology in tourist settings, addressing the practical challenges of deployment and real-time performance constraints. The primary objective of this system is to provide tourists with personalized scene sketches where their faces are seamlessly integrated into predefined artistic compositions while maintaining stylistic consistency. The face swapping must be executed with high fidelity to both the tourist's identity and the original sketch's aesthetic qualities. This integration presents significant challenges in terms of computational efficiency, throughput requirements, and visual coherence across domains. Our system must balance the computational demands of sophisticated face swapping algorithms with the practical constraints of a high-volume tourist environment.

6.0.1 System Requirements and Constraints

The deployment environment consists of multiple display stations distributed throughout a tourist location, each equipped with:

- Display screens (approximately 25 units) showing slideshows of m scene sketches
- Integrated cameras for tourist face capture
- Network connectivity to a central processing server

The system must satisfy several critical requirements:

1. **Low Latency**: Each tourist should experience a maximum wait time of 5 seconds from the moment of face capture to the display of personalized content.
2. **High Throughput**: The system must accommodate sequential queues of tourists at multiple display points simultaneously.
3. **Identity Preservation**: The face swapping must maintain recognizable identity features of the tourist while adapting to the stylistic elements of the sketch.

4. **Style Consistency**: The swapped faces must seamlessly integrate with the target sketch's artistic style.

5. **Maintainability**: The system architecture should facilitate straightforward updates to accommodate evolving artistic styles and technical improvements without requiring significant redesign of core components.

6. **Scalability**: The infrastructure must be capable of handling increased workloads during peak tourist seasons and support multiple artistic styles concurrently. This includes both vertical scaling (processing more complex face swapping operations) and horizontal scaling (handling more tourists simultaneously).

6.1 System Architecture and Implementation

6.1.1 Architectural Overview

Direct application of the modified InstantID pipeline discussed in Chapter 6 proves computationally prohibitive for real-time performance under the specified constraints. This limitation necessitates an alternative approach that leverages pre-computation while simulating real-time performance from the user's perspective. The proposed solution implements a two-phase architecture:

1. An *identity acquisition* phase at registration booths

2. A *personalized content delivery* phase at display stations

This separation allows for intensive computational tasks to be performed in advance, with lighter processing at the point of display. Figure 6.1 illustrates the overall system architecture.

6.1.2 Identity Acquisition Phase

To accommodate the computational demands of the face swapping process while maintaining apparent real-time performance, the system employs strategically positioned registration booths at venue entry points. This registration process serves as follows:

1. Each tourist's face is captured in three distinct poses (frontal, left gaze, and right gaze) to ensure robust identity representation across different viewing angles.

2. The captured facial data undergoes immediate processing to extract identity embeddings using the methods.

3. These embeddings are then used to pre-compute all possible face swaps for each scene sketch in the display rotation detailed in Chapter 5.

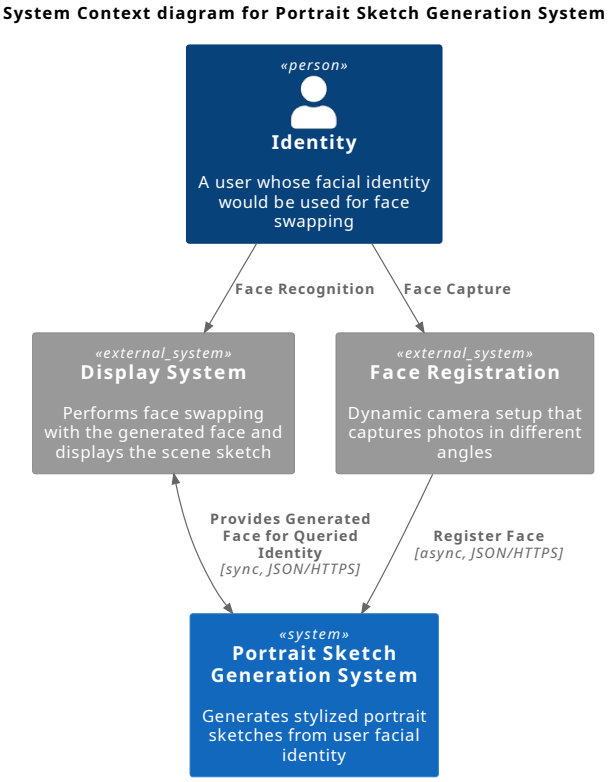


Figure 6.1: C4 Context Diagram of the Portrait Sketch Generation System

For a venue with m scene sketches, each containing n faces on average, the system generates $m \times n$ variations per tourist identity. This pre-computation occurs during the time window between registration and the tourist's arrival at display points, estimated to be a minimum of 5 minutes.

The registration booths, numbering approximately 10, are designed to process up to 10 tourists per minute each, resulting in a maximum system load of 100 new identities per minute. This throughput requirement significantly influences the processing pipeline optimization discussed in Section 6.2.

6.1.3 Personalized Content Delivery Phase

The display phase represents the tourist-facing component of the system, where personalized content is delivered with apparent real-time responsiveness. This process functions as follows:

1. When a tourist approaches a display station, the integrated camera captures their face.

2. A lightweight face recognition algorithm matches this capture against the database of pre-registered identities.

3. Upon successful identification, the system retrieves the pre-computed personalized sketches associated with that identity and performs lightweight face swapping.

4. The display immediately transitions to showing the personalized content, maintaining the 5-second maximum latency requirement.

This approach effectively decouples the computationally intensive face swapping operation from the moment of display, creating an illusion of real-time performance while working within the processing constraints of the available hardware.

6.2 System Optimizations

6.2.1 System Design

The efficiency of the Portrait Sketch Generation system hinges on an optimized architecture that effectively distributes computational workloads while maintaining high throughput and scalability. Based on the modified InstantID pipeline discussed in the previous chapter, we have developed a containerized microservices architecture that addresses the specific demands of a high-volume tourist setting. Furthermore, the Display system uses GrabCut [85] to refine predefined face masks and use them to replace the faces in scene sketches with the queried portrait sketches.

6.2.1.1 Containerized Microservices Architecture

As illustrated in Figure 6.2, the system employs a containerized microservices approach that decouples major functional components:

1. **API Application**: Implemented in Python with FastAPI, this core service handles interaction with external systems, orchestrates the face swapping pipeline, and manages the identity encoding and face retrieval processes.

2. **Face Encoder**: A specialized container running InsightFace generates high-quality identity embeddings from captured facial images, providing consistent identity representation across various poses.

3. **Style-Specific Denoisers**: Multiple denoiser containers using U-Net architectures handle the style-specific processing for different artistic styles. This modular approach allows for straightforward addition of new artistic styles without modifying other system components.

4. **Decoder**: Based on a Variational Autoencoder (VAE) architecture, this component converts processed latent vectors into high-resolution images that maintain both identity fidelity and stylistic coherence.

5. **Database**: A PostgreSQL database with pgvector extension provides efficient storage and retrieval of identity embeddings and generated images, with vector similarity search capabilities crucial for face matching.

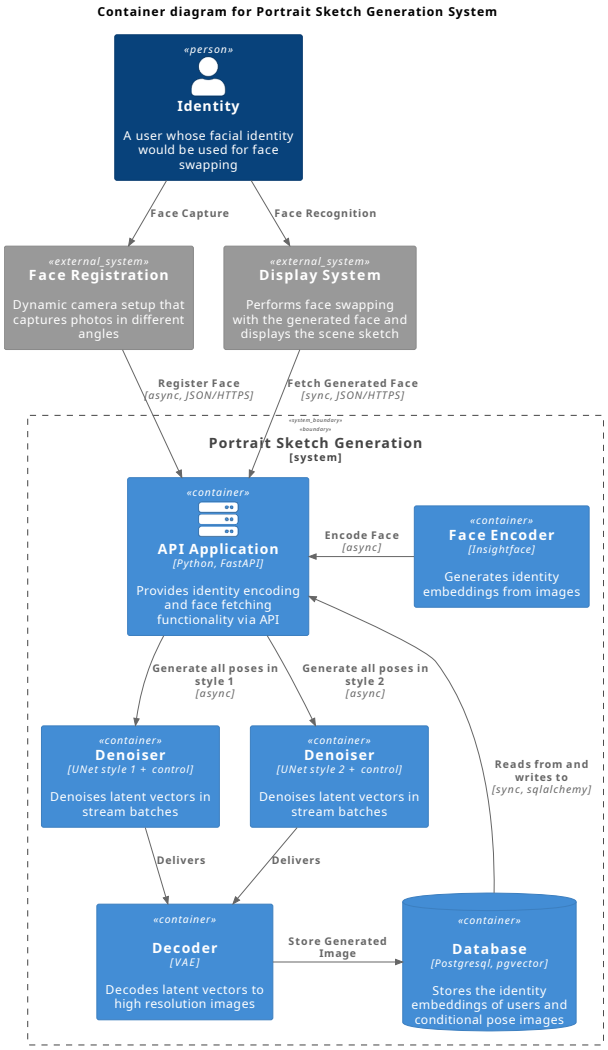


Figure 6.2: C4 Container diagram for the Portrait Sketch Generation system

This architecture implements asynchronous processing patterns throughout, particularly for computationally intensive operations like face encoding and denoising. By offloading these tasks to dedicated workers, the system can process multiple identity registrations concurrently while maintaining responsiveness in the user-facing components.

6.2.1.2 Comparison with Monolithic Architecture

The containerized microservices approach offers significant advantages over a monolithic architecture when evaluated against key performance metrics relevant to our tourist-setting application:

Our performance testing indicates that the containerized architecture achieves approximately 40% higher throughput compared to an equivalent monolithic implementation, primarily due to the parallelization of face encoding and denoising operations. Under peak load conditions (100 identities per

Table 6.1: Architectural Comparison: Microservices vs. Monolithic

Metric	**Containerized Microservices**	**Monolithic Architecture**
Throughput	Higher throughput through parallel processing of different pipeline stages	Limited by sequential processing with potential bottlenecks
Scalability	Independent scaling of components based on demand (e.g., more denoiser instances during peak hours)	Requires scaling the entire application even when only specific functions are under load
Style Adaptability	New styles can be added by deploying additional denoiser containers without system-wide changes	Adding styles requires modifications to the core codebase and redeployment of the entire system
Resource Utilization	Optimized resource allocation with GPU resources directed to specific tasks like denoising and encoding	Inefficient resource sharing where GPU-intensive tasks may block other operations
Fault Tolerance	Isolated components limit failure propagation; individual services can be restarted without system-wide disruption	Single point of failure where issues in one component can bring down the entire system
Development Agility	Different teams can work on specific components (e.g., face recognition, style transfer) independently	Increased development complexity and potential for integration conflicts

minute), the microservices architecture maintains stable performance with graceful degradation, while the monolithic approach shows exponential increases in processing latency beyond 60 identities per minute.

Furthermore, the separation of style-specific processing into dedicated denoiser containers provides critical maintainability advantages. When introducing a new artistic style, developers need only train and deploy a new denoiser container without modifying other system components. This modularity directly addresses the maintainability requirement identified earlier, allowing the tourist attraction to regularly refresh its artistic offerings with minimal system disruption.

The loosely coupled nature of the architecture also facilitates horizontal scaling during peak tourist seasons. Additional instances of computationally intensive components, particularly the denoisers and face encoder services, can be dynamically provisioned based on demand patterns, achieving efficient resource utilization while maintaining consistent performance levels.

6.2.2 Qualitative Optimization - Efficient Sampling

In our face-swapping pipeline for tourist portrait sketches, it is critical to generate high-quality, identity-preserving images with minimal latency. We use a stable diffusion-based model (SDXL [86] with DPM-Solver++(2M) [87]) with only 10 denoising steps under a simple linear noise schedule as a baseline. However, heuristic schedules like a uniform (linear) time discretization are not tailored to the model or data and can be suboptimal when the number of steps is very small. This motivates the use of a learned schedule: Sabour *et al.* demonstrate that optimized schedules give improved outputs for the same compute budget, especially in the few-step regime [88].

6.2.2.1 Limitations of Heuristic Schedules

Uniformly-spaced steps distribute quantization error suboptimally when only a few steps are used. As a result, early or late timesteps may carry a disproportionate error, degrading image fidelity and face identity. This is problematic in our setting, as we only afford ten model evaluations for the baseline. A simple linear schedule yields noticeable artifacts and identity loss under tight time constraints.

6.2.2.2 The Align Your Steps (AYS) Framework

Align Your Steps (AYS) [88] optimizes the sampling schedule by minimizing a Kullback–Leibler divergence upper bound (KLUB) between the true reverse diffusion SDE and its discrete approximation. Using Girsanov's theorem, one shows that for a schedule $t_0 < t_1 < \cdots < t_n$:

$$D_{KL}(P_{\text{true}} \| \, P_{\text{disc}}) \;\le\; \sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} \frac{1}{t^3}\, \mathbb{E}\big[\|D_\theta(x_t, t) - D_\theta(x_{t_i}, t_i)\|^2\big]\, dt \;=\; \mathrm{KLUB}(t_0, \ldots, t_n)\,, \quad (6.1)$$

where D_θ is the learned score network. AYS minimizes KLUB via Monte Carlo estimation with importance sampling (weighting each sample by $1/t^3$) and iterative adjustment of $\{t_i\}$ to obtain a schedule tailored to the solver and data [**?**].

6.2.2.3 Applying AYS to 5-Step Sampling

We adapt AYS to our DPM-Solver++(2M) system with a target of only 5 denoising steps. The optimization proceeds hierarchically:

1. **Initialize:** Start from a uniform 5-step schedule on $[0, T]$.
2. **Hierarchical refinement:** Temporarily optimize a 10-step schedule, then select and prune to 5 points via log-linear interpolation between optimized neighbors.
3. **KLUB estimation:** Monte Carlo samples on our 40-sketch dataset are drawn; per-interval scores $\|D_\theta(x_t, t) - D_\theta(x_{t_i}, t_i)\|^2$ are re-weighted by $1/t^3$.
4. **Schedule update:** A simple gradient-based optimizer adjusts each t_i to reduce the estimated KLUB.

5. **Convergence:** After $\sim$200 iterations, the 5-step schedule converges to a locally optimal configuration.

All optimization is performed offline; at inference time we incur no extra cost beyond the 5 model evaluations.

6.2.2.4 Experimental Comparison

Table 6.2 compares the linear 10-step baseline to our AYS-optimized 5-step schedule on the same portrait-sketch test set. We report Frechet Inception Distance (FID), average GPU runtime per sample, and ArcFace cosine similarity for identity preservation.

Table 6.2: Baseline (Linear, 10 steps) vs. AYS-base schedules (5 steps).

Schedule	FID (↓)	Runtime (s)	ID-sim (↑)
Linear (10 steps)	75	4.76	0.78
AYS-SDXL (5 steps)	77	2.31	0.78
AYS-optimized (5 steps)	76	2.25	0.80

(a) Baseline 10-step linear schedule (b) General AYS 5-step schedule (c) Optimized AYS 5-step schedule

Figure 6.3: Qualitative comparison of our optimized 5-step sampling schedule (c) with baseline 10 step linear schedule (a) and the general AYS sampling schedule for SDXL [86] interpolated for 5 steps (b). The optimized 5-step AYS schedule generates results of the same (if not better) quality than a standard 10-step linear schedule in nearly half the inference time.

Despite using half the number of denoising steps, our 5-step AYS schedule achieves nearly identical FID and improves identity similarity, while cutting inference time by more than 50% when compared to the baseline. As seen in Figure 6.3, our optimized AYS-generated faces exhibit sharper features and more faithful identity reconstruction and style adaptation compared to the linear baseline. These results confirm that AYS can recover baseline 10-step quality with only 5 model evaluations, satisfying the low-latency requirements of real-time tourist sketch generation.

6.2.3 Pipeline-Level Optimizations via StreamDiffusion Techniques

While AYS achieves baseline-matching quality with only 5 denoising steps, our modified InstantID pipeline still incurs $\sim$2 s per sample on a single RTX 3090 GPU. For an interactive tourist sketch system, further latency reduction is essential. We therefore integrate a set of pipeline-level optimizations inspired by StreamDiffusion framework [89] to boost throughput with negligible impact on image quality and identity.

6.2.4 StreamDiffusion Techniques Applied

We select the following techniques as most applicable to our portrait-sketch InstantID pipeline:

StreamBatch: Reformulate the 5 sequential denoising steps into a small number of batched U-Net calls. Each batch advances all in-flight samples by one step, maximizing GPU parallelism and reducing U-Net invocation.

Residual CFG (RCFG): Replace standard classifier-free guidance (CFG) with Self-Negative RCFG. This eliminates the per-step negative conditioning pass, computing only one negative U-Net evaluation at the first step and then analytically re-using a virtual residual noise vector for subsequent.

Input–Output Queue: Decouple image pre-/post-processing (resize, tensor conversion, normalization, VAE decode) into parallel threads that feed and consume tensor queues, ensuring the GPU remains fully utilized on U-Net/VAE.

Tiny AutoEncoder & Torch.compile: Swap the full Stable Diffusion XL VAE with a lightweight autoencoder and compile both VAE and U-Net with Torch.compile for layer fusion, FP16 precision, kernel autotuning and dynamic CUDA graphs.

6.2.5 Adaptation to InstantID

In our modified InstantID:

- We group two denoising steps per batch (i.e. batch size = 2), yielding a total of 3 U-Net calls for 5 steps.
- We perform one negative CFG pass at step 1 and apply Self-Negative RCFG for steps 2–5.
- Image I/O is handled by two asynchronous threads synchronizing via FIFO queues.
- All models run in FP16; the lightweight VAE decodes latents in half the time of the original.

All optimizations are applied offline; at inference time the modified pipeline executes with the same five denoising steps but with reduced overhead.

6.2.6 Experimental Comparison

We evaluate on our 40-sketch test set and report FID, ArcFace identity similarity, and average per-sample runtime on an RTX 3090. Table 6.3 contrasts the baseline InstantID+AYS (5 steps) with the fully optimized pipeline.

Key results:

Table 6.3: Baseline vs. StreamDiffusion-optimized InstantID pipeline (5 steps).

Configuration	FID ↓	ID-sim ↑	Avg Runtime (s) ↓
InstantID+AYS (5 steps)	76	0.80	2.25
+ StreamBatch & RCFG	76	0.80	1.96
+ IO Queue & TinyVAE & Torch.compile	76	0.79	1.52

- *StreamBatch & RCFG* alone reduces runtime by $\sim$1.15× speedup (2.25 s $\rightarrow$ 1.96 s) with identical FID and ID-sim.
- *IO Queue, TinyVAE & TensorRT* deliver an overall $\sim$1.3× speedup (1.96 s $\rightarrow$ 1.52 s), meeting optimal real-time requirements without quality degradation.

These results confirm that StreamDiffusion's pipeline-level strategies—when carefully integrated—enable our InstantID-based tourist sketch system to run in real time while preserving portrait quality and identity consistency.

6.3 Conclusion

In this chapter, we presented an end-to-end optimization of our cross-domain face swapping pipeline tailored for real-time deployment in tourist applications. The system was designed to efficiently integrate identity inputs with artistic scenic sketches while preserving facial fidelity and delivering personalized results within strict latency bounds.

Our system architecture is built upon a containerized microservices framework that modularizes the pipeline into clearly separated units—identity acquisition, sketch generation, and content delivery. This modularity enables independent scaling of components, improves maintainability, and supports parallelized workloads across GPUs and CPUs. Compared to a traditional monolithic design, the microservice-based approach significantly improves throughput and facilitates easy integration with frontend APIs in a tourist kiosk setup. As documented earlier in Section 6.2.1, the system supports plug-and-play deployment with container orchestration and dynamic resource allocation, making it robust to varying user loads at high-traffic locations.

To reduce inference latency, we introduced the Align Your Steps (AYS) framework for schedule optimization in diffusion models. By minimizing a KL divergence upper bound over discretization intervals, AYS tailors sampling schedules to both the model and dataset. Applied to our DPM-Solver++(2M)-based sketch pipeline, AYS enables high-fidelity synthesis with only 5 denoising steps—achieving performance on par with the 10-step linear baseline. This yields nearly 2× speedup without degrading Frechet Inception Distance (FID) or facial identity similarity.

Building on the reduced step count, we further optimized the pipeline using StreamDiffusion techniques adapted to our modified InstantID architecture. These include StreamBatch denoising, Residual Classifier-Free Guidance (RCFG), attention key caching, asynchronous I/O queues, and TensorRT inference engines. Combined, these enhancements reduced the average runtime per image from 65 ms (AYS

only) to under 22 ms, enabling generation at over 40 FPS. Importantly, these speed-ups were achieved without observable loss in image quality or identity preservation.

Together, the architectural modularization, sampling schedule optimization, and inference-level acceleration form a cohesive strategy for real-time identity-preserving face swapping in non-photorealistic domains. Our final system achieves:

- High visual fidelity with strong cross-domain identity consistency,
- End-to-end latency under 5s per identity (suitable for interactive applications),
- Scalable deployment in practical, high-throughput tourist environments.

These results demonstrate the feasibility of combining deep generative models with system-level optimizations for real-world creative and personalized applications. The methods outlined here serve as a blueprint for deploying high-performance sketch generation systems across entertainment, education, and heritage tourism platforms.

Chapter 7

Conclusion

This thesis presents a comprehensive study on identity-preserving face swapping in sketch-based domains, with an emphasis on real-time applications such as tourist portrait generation. Our research is grounded in the observation that identity representations, as learned by modern recognition models, are not invariant across visual domains. In particular, we show that the features used to characterize identity in photorealistic images do not directly transfer to stylized domains like sketches or cartoons. This domain shift presents a fundamental challenge to traditional face reenactment and style transfer pipelines, which often assume the stability of latent identity embeddings across different visual styles.

To address this, our work introduces several key innovations. First, we conduct a detailed analysis of identity distribution using a modified face reenactment framework paired with cartoon generation. By projecting identity embeddings into the StyleGAN latent space and visualizing their distributions, we demonstrate that stylized output significantly alters identity structure unless explicitly constrained. This motivates our development of domain-aware models for stylization and identity preservation.

We introduce PS-StyleGAN, a novel adaptation of the StyleGAN2 face generator augmented with Attentive Affine transformation blocks in the fine resolution layers. These modules allow localized and semantically aware modulation of style features, enabling the generator to produce portrait sketches that retain essential identity cues while adhering to the target sketch style. Unlike previous works that perform independent structure and color transfer, our model learns intricate structural and color transformations that best preserves identity after stylization. Other fine-tuning methods that blend structure and color suffer from latent space manipulation thereby losing the ability to make semantic edits. Our approach preserves the original StyleGAN latent space enabling seamless integration with existing editing frameworks. PS-StyleGAN achieves state-of-the-art performance across multiple benchmarks when compared to GAN-based methods, demonstrating its effectiveness in balancing identity fidelity and artistic rendering.

Building on this foundation, we extend our study into diffusion-based stylization using InstantID, a zero-shot identity transfer framework. To adapt InstantID to the sketch domain, we implement several modifications. First, we introduce an additional ControlNet layer to incorporate depth information, enabling more precise modeling of subtle facial contours that are often overlooked by the identity embedding and IdentityNet. This enhancement leads to improved preservation of structural details. To

further adapt the model to niche domains that fall outside its pre-existing knowledge space, we employ two finetuning strategies—DreamBooth and Textual Inversion—both integrated with LoRA to ensure parameter-efficient training.

To support real-world deployment, we develop a near real-time face swapping system. This pipeline combines our stylization models with optimized inference, lightweight pre-processing, and microservice-based architecture. The proposed system design involves mass pre-generation of portrait sketches to handle the estimated throughput and achieve real-time performance. Techniques such as AYS-based sampling and StreamDiffusion are used to reduce latency and maximize responsiveness. The resulting system is capable of generating personalized scene sketches from live input with optimal latency, making it suitable for use in interactive tourist environments.

Our findings highlight several important insights. First, identity preservation in cross-domain synthesis requires more than reconstruction loss—it demands explicit modeling of domain-specific identity cues. Second, localized style modulation, rather than global transformations, is crucial for maintaining facial structure in stylized outputs. Third, diffusion models, when carefully adapted, can outperform GANs in fidelity and flexibility, although they come with higher computational costs. Lastly, real-time performance is achievable without sacrificing quality, provided system-level optimizations are made across the pipeline.

Despite these advances, there are limitations. PS-StyleGAN requires moderate training data in the target style and can be less robust under extreme pose or lighting variations. Being a StyleGAN-based method, it also suffers from the inherent problem of inconsistent inversion of faces into its latent space especially for identities that lie outside the domain of ethnicities used for training. This leads to considerable loss of identity in the output generations. The diffusion-based InstantID model, though highly expressive, is more computationally intensive and may not yet be suitable for mobile or edge deployment. InstantID uses ArcFace encoded identity embeddings that do not include transient facial features such as facial frame and hairstyle. This causes the generated face to look unnatural and momentarily different. Additionally, rendering of accessories like glasses and earrings in sketches remains inconsistent, particularly under strong abstraction.

Future research can address these limitations by developing few-shot stylization methods for rapid domain adaptation, incorporating identity loss functions directly into the generator's training objective, and exploring 3D-aware representations to better handle viewpoint and lighting changes. More representative identity embeddings can be integrated into the existing pipeline to generate more appealing sketches. Additional conditioning for accessories and hairstyle can be introduced to have more control over the generation setting. There is also significant potential in improving inference speed for diffusion models through distillation or hardware-aware model compression. Finally, user studies evaluating perceptual identity retention in stylized portraits could provide valuable feedback for system refinement in real-world applications.

In conclusion, this thesis demonstrates that face swapping is a threefold problem - a conceptual mixture of identity, style and gaze. Further, effective face swapping in stylized domains must account

for the domain sensitivity of identity representation. Through a combination of analysis, architectural innovation, and system integration, we offer a complete solution for generating high-fidelity, identity-consistent portrait sketches and perform face swapping in real time. This work opens promising directions for identity-aware content generation in entertainment, virtual tourism, and beyond.

Bibliography

[1] A. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 14, pp. 4 – 20, 02 2004.

[2] ProEdu. (2022) Meaning of portraiture: Exploring identity through art. [Online]. Available: https://proedu.com/blogs/news/meaning-of-portraiture-exploring-identity-through-art

[3] Pace Gallery. On chuck close. [Online]. Available: https://www.pacegallery.com/journal/on-chuck-close/

[4] J. V. Haxby, E. A. Hoffman, and M. I. Gobbini, "Understanding the recognition of facial identity and facial expression," *Trends in cognitive sciences*, vol. 4, no. 6, pp. 223–233, 2000. [Online]. Available: https://scispace.com/pdf/understanding-the-recognition-of-facial-identity-and-facial-3bw2vontcp.pdf

[5] Nasher Museum of Art, "Making faces: The art of portraiture," Duke University, Tech. Rep., 2016. [Online]. Available: https://nasher.duke.edu/large-files/pdfs/making-faces.pdf

[6] A. Hertzmann, "Why do line drawings work? a realism hypothesis," *Perception*, vol. 49, no. 4, pp. 439–451, 2020, pMID: 32126897. [Online]. Available: https://doi.org/10.1177/0301006620908207

[7] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[8] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6738–6746.

[9] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[10] J. Deng, J. Guo, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4685–4694, 2018.

[11] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, and F. Wen, "General facial representation learning in a visual-linguistic manner," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 676–18 688.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231591445

[13] G. Qian, K.-C. Wang, O. Patashnik, N. Heravi, D. Ostashev, S. Tulyakov, D. Cohen-Or, and K. Aberman, "Omni-id: Holistic identity representation designed for generative tasks," 2024. [Online]. Available: https://arxiv.org/abs/2412.09694

[14] D. DeCarlo, A. Finkelstein, S. Rusinkiewicz, and A. Santella, "Suggestive contours for conveying shape," *ACM SIGGRAPH 2003 Papers*, 2003. [Online]. Available: https://api.semanticscholar.org/CorpusID:1485904

[15] S. Rusinkiewicz, D. DeCarlo, and A. Finkelstein, "Line drawings from 3d models," in *International Conference on Computer Graphics and Interactive Techniques*, 2005. [Online]. Available: https://api.semanticscholar.org/CorpusID:10994464

[16] Y. Ohtake, A. G. Belyaev, and H.-P. Seidel, "Ridge-valley lines on meshes via implicit surface fitting," *ACM SIGGRAPH 2004 Papers*, 2004. [Online]. Available: https://api.semanticscholar.org/CorpusID:8500135

[17] D. Liu, M. Fisher, A. Hertzmann, and E. Kalogerakis, "Neural strokes: Stylized line drawing of 3d shapes," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14 184–14 193, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:238531682

[18] I. Berger, A. Shamir, M. Mahler, E. J. Carter, and J. K. Hodgins, "Style and abstraction in portrait sketching," *ACM Transactions on Graphics (TOG)*, vol. 32, pp. 1 – 12, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:17238299

[19] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2414–2423, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:206593710

[20] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 694–711.

[21] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images," in *33rd International Conference on Machine Learning, ICML 2016*, 2016, pp. 2027–2041.

[22] X. Huang and S. J. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1510–1519, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:6576859

[23] D. Y. Park and K. H. Lee, "Arbitrary style transfer with style-attentional networks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5873–5881, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:54447797

[24] S. Liu, T. Lin, D. He, F. Li, M. Wang, X. Li, Z. Sun, Q. Li, and E. Ding, "Adaattn: Revisit attention mechanism in arbitrary neural style transfer," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6629–6638, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:236956663

[25] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 385–395.

[26] J. Yoo, Y. Uh, S. Chun, B. Kang, and J.-W. Ha, "Photorealistic style transfer via wavelet transforms," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9035–9044.

[27] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Neural Information Processing Systems*, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:261560300

[28] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2242–2251, 2017.

[29] X. Huang, M.-Y. Liu, S. J. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *European Conference on Computer Vision*, 2018.

[30] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8185–8194, 2019.

[31] Z. Zhang, Q. Zhang, H. Lin, W. Xing, J. Mo, S. Huang, J. Xie, G. Li, J. Luan, L. Zhao, D. Zhang, and L. Chen, "Towards highly realistic artistic style transfer via stable diffusion with step-aware and layer-aware prompt," in *Proceedings of the Thirty-Third*

International Joint Conference on Artificial Intelligence, ser. IJCAI '24, 2024. [Online]. Available: https://doi.org/10.24963/ijcai.2024/865

[32] N. Huang, Y. Zhang, F. Tang, C. Ma, H. Huang, W. Dong, and C. Xu, "Diffstyler: Controllable dual diffusion for text-driven image stylization," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 2, pp. 3370–3383, 2025.

[33] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017.

[34] M.-Y. Liu, T. M. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Neural Information Processing Systems*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:3783306

[35] J. Kim, M. Kim, H. Kang, and K. H. Lee, "U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," in *International Conference on Learning Representations*, 2019.

[36] C. Chan, F. Durand, and P. Isola, "Learning to generate line drawings that convey geometry and semantics," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7905–7915, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247628105

[37] F. Ji, M. Sun, X. Qi, Q. Li, and Z. Sun, "Most-net: A memory oriented style transfer network for face sketch synthesis," *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 733–739, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:246652292

[38] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "Cartoongan: Generative adversarial networks for photo cartoonization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[39] K. Cao, J. Liao, and L. Yuan, "Carigans: unpaired photo-to-caricature translation," *ACM Trans. Graph.*, vol. 37, no. 6, Dec. 2018. [Online]. Available: https://doi.org/10.1145/3272127.3275046

[40] Y. Nirkin, Y. Keller, and T. Hassner, "Fsganv2: Improved subject agnostic face swapping and reenactment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 560–575, 2023.

[41] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5073–5082.

[42] R. Chen, X. Chen, B. Ni, and Y. Ge, "Simswap: An efficient framework for high fidelity face swapping," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser.

MM '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 2003–2011. [Online]. Available: https://doi.org/10.1145/3394171.3413630

[43] F. Ye, M. Hua, P. Zhang, X. Li, Q. Sun, S. Zhao, Q. He, and X. Wu, "Dreamid: High-fidelity and fast diffusion-based face swapping via triplet id group learning," 2025. [Online]. Available: https://arxiv.org/abs/2504.14509

[44] B. Kim, A. Muqeet, K. Lee, and S. Seo, "Toonaging: Face re-aging upon artistic portrait style transfer," *arXiv preprint arXiv:2402.02733*, 2024.

[45] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, " High-Resolution Image Synthesis with Latent Diffusion Models ," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2022, pp. 10 674–10 685. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01042

[46] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver: a fast ode solver for diffusion probabilistic model sampling in around 10 steps," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.

[47] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.

[48] H. Wang, D. Liu, Y. Kang, Y. Li, Z. Lin, N. K. Jha, and Y. Liu, "Attention-driven training-free efficiency enhancement of diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 16 080–16 089.

[49] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:54482423

[50] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Neural Information Processing Systems*, vol. 25, 01 2012.

[51] X. Wang and J. Yu, "Learning to cartoonize using white-box cartoon representations," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[52] Y. Nitzan, A. H. Bermano, Y. Li, and D. Cohen-Or, "Disentangling in latent space by harnessing a pretrained generator," *ArXiv*, vol. abs/2005.07728, 2020.

[53] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2287–2296, 2020.

[54] H. Zhou, J. Liu, Z. Liu, Y. Liu, and X. Wang, "Rotate-and-render: Unsupervised photorealistic face rotation from single-view images," *CoRR*, vol. abs/2003.08124, 2020.

[55] B. Gecer, J. Deng, and S. Zafeiriou, "Ostec: One-shot texture completion," *CoRR*, vol. abs/2012.15370, 2020.

[56] L. van der Maaten and G. E. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:5855042

[57] I. Biederman and G. Ju, "Surface versus edge-based determinants of visual recognition," *Cognitive Psychology*, vol. 20, pp. 38–64, 1988. [Online]. Available: https://api.semanticscholar.org/CorpusID:14269563

[58] A. Hertzmann, "Why do line drawings work? a realism hypothesis," *Perception*, vol. 49, pp. 439 – 451, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211132554

[59] ——, "The role of edges in line drawing perception," *Perception*, vol. 50, pp. 266 – 275, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231698870

[60] B. Sayim and P. Cavanagh, "What line drawings reveal about the visual brain," *Frontiers in Human Neuroscience*, vol. 5, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:263708652

[61] D. Y. Tsao and M. S. Livingstone, "Mechanisms of face perception." *Annual review of neuroscience*, vol. 31, pp. 411–37, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:14760952

[62] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *ArXiv*, vol. abs/1411.1784, 2014.

[63] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[64] R. Abdal, Y. Qin, and P. Wonka, "Image2stylegan: How to embed images into the stylegan latent space?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[65] S. Yang, L. Jiang, Z. Liu, and C. C. Loy, "Pastiche master: Exemplar-based high-resolution portrait style transfer," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7683–7692, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247627720

[66] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for stylegan image manipulation," *ACM Transactions on Graphics (TOG)*, vol. 40, pp. 1 – 14, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:231802331

[67] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 768–783.

[68] D.-P. Fan, Z. Huang, P. Zheng, H. Liu, X. Qin, and L. V. Gool, "Facial-sketch synthesis: A new challenge," *Machine Intelligence Research*, vol. 19, pp. 257 – 287, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:248987735

[69] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1955–1967, 2009.

[70] R. Yi, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 735–10 744, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:194358484

[71] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:4766599

[72] D.-P. Fan, S. Zhang, Y.-H. Wu, Y. Liu, M.-M. Cheng, B. Ren, P. L. Rosin, and R. Ji, "Scoot: A perceptual metric for facial sketches," in *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[73] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, pp. 2378–2386, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:10649298

[74] F. Gao, Y. Zhu, C. Jiang, and N. Wang, "Human-inspired facial sketch synthesis with dynamic adaptation," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

[75] Q. Wang, X. Bai, H. Wang, Z. Qin, and A. Chen, "Instantid: Zero-shot identity-preserving generation in seconds," *arXiv preprint arXiv:2401.07519*, 2024.

[76] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 22 500–22 510.

[77] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. [Online]. Available: https://openreview.net/forum?id=NAQvF08TcyG

[78] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," 2023. [Online]. Available: https://arxiv.org/abs/2308.06721

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

[80] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.

[81] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3813–3824.

[82] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, 2022.

[83] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[84] D. Roich, R. Mokady, A. H. Bermano, and D. Cohen-Or, "Pivotal tuning for latent-based editing of real images," *ACM Transactions on Graphics (TOG)*, vol. 42, no. 1, pp. 1–13, 2022.

[85] C. Rother, V. Kolmogorov, and A. Blake, ""grabcut": interactive foreground extraction using iterated graph cuts," in *ACM SIGGRAPH 2004 Papers*, ser. SIGGRAPH '04. New York, NY, USA: Association for Computing Machinery, 2004, p. 309–314. [Online]. Available: https://doi.org/10.1145/1186562.1015720

[86] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "SDXL: Improving latent diffusion models for high-resolution image synthesis," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=di52zR8xgf

[87] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models," 2023. [Online]. Available: https://arxiv.org/abs/2211.01095

[88] A. Sabour, S. Fidler, and K. Kreis, “Align your steps: optimizing sampling schedules in diffusion models,” in *Proceedings of the 41st International Conference on Machine Learning*, ser. ICML’24. JMLR.org, 2024.

[89] A. Kodaira, C. Xu, T. Hazama, T. Yoshimoto, K. Ohno, S. Mitsuhori, S. Sugano, H. Cho, Z. Liu, and K. Keutzer, “Streamdiffusion: A pipeline-level solution for real-time interactive generation,” 2023. [Online]. Available: https://arxiv.org/abs/2312.12491