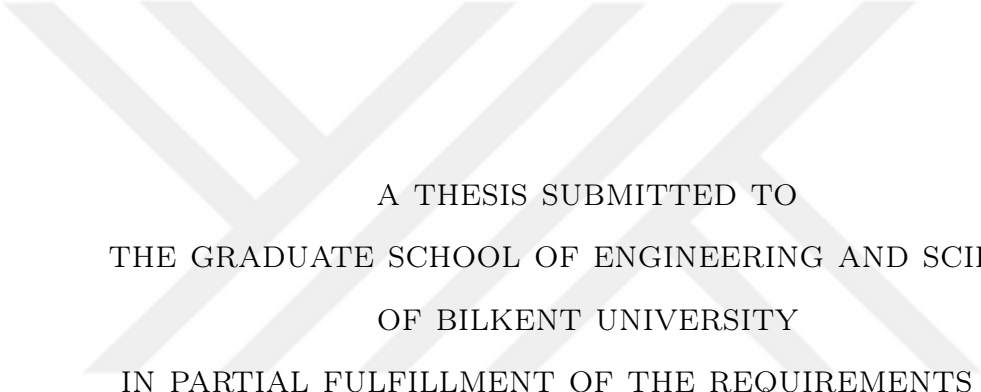


MIXGAN: DUAL PATH STYLEGAN FUSION FOR DIVERSE AND EDITABLE INPAINTING



A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

By
Mustafa Utku Aydoğdu
September 2025

MixGAN: Dual Path StyleGAN Fusion for Diverse and Editable
Inpainting

By Mustafa Utku Aydoğdu

September 2025

We certify that we have read this thesis and that in our opinion it is fully adequate,
in scope and in quality, as a thesis for the degree of Master of Science.

Ayşegül Dünder Boral(Advisor)

Uğur Gündükbay

Ramazan Gökberk Cinbiş

Approved for the Graduate School of Engineering and Science:

Orhan Arıkan
Director of the Graduate School

ABSTRACT

MIXGAN: DUAL PATH STYLEGAN FUSION FOR DIVERSE AND EDITABLE INPAINTING

Mustafa Utku Aydoğdu

M.S. in Computer Engineering

Advisor: Ayşegül Dünder Boral

September 2025

While generative models have advanced image inpainting, achieving diverse, editable results that blend seamlessly remains a key challenge, particularly for GANs, which struggle to balance reconstruction fidelity with editability. This thesis introduces a novel framework to address these gaps, unifying diversity, fidelity, and user control within a single GAN-based system. First, a new architecture leverages StyleGAN’s feature space by employing parallel generation pathways—one from the encoded image and another from a random latent sample. Through novel **mixing networks with dual feedback**, these pathways are dynamically aligned in the feature space, rather than the latent space, enabling diverse, high-quality inpaintings that maintain structural integrity. Next, the framework perfects the fidelity-editability balance using **masked skip connections** with an **information bottleneck**. This forces the model to rely on editable low-rate codes for missing regions while using high-rate features to flawlessly reconstruct visible parts. The result is seamless boundary transitions and preserved control for attribute-specific edits, overcoming common artifacts in prior work. Together, these contributions provide a robust solution that advances GAN-based inpainting. The method generates diverse, high-fidelity, and editable content by operating in the feature space and managing information flow. Extensive experiments demonstrate that our framework surpasses existing approaches in both quality and diversity, offering a fast, controllable system for image restoration and editing.

Keywords: Generative Adversarial Networks (GANs) , Image Inpainting, GAN Inversion, Style-GAN, Generative Models, Image Editing, Diverse Generation.

ÖZET

MIXGAN: ÇEŞİTLİ VE DÜZENLENEBİLİR İÇBOYAMA İÇİN ÇİFT YOLLU STYLEGAN FÜZYONU

Mustafa Utku Aydoğdu

Bilgisayar Mühendisliği, Yüksek Lisans

Tez Danışmanı: Ayşegül Dünder Boral

September 2025

Üretken modeller görüntü içboyama alanını ileriye taşımış olsa da, özellikle yeniden yapılandırma doğruluğu ile düzenlenebilirlik arasında denge kurmakta zorlanan üretken çekişmeli ağlar için, mevcut içerikle sorunsuz bütünleşen çeşitli ve düzenlenebilir sonuçlar elde etmek önemli bir zorluktur. Bu çalışma, tek bir üretken çekişmeli ağ tabanlı sistem içinde çeşitliliği, doğruluğu ve kullanıcı kontrolünü birleştirerek bu eksiklikleri gidermek için özgün bir yöntem sunmaktadır. İlk olarak, yeni bir mimari, biri kodlanmış görüntüden diğeri ise rastgele bir gizil örneklemden gelen paralel üretim yolları kullanarak Style-GAN'ın öznitelik uzayından faydalanır. **Çift yönlü geri beslemeli özgün karıştırma ağları** sayesinde bu yollar, gizil uzay yerine öznitelik uzayında dinamik olarak hizalanır. Bu da yapısal bütünlüğü koruyan, çeşitli ve yüksek kaliteli içboyamalar sağlar. Ardından, önerilen yöntem, **bilgi darboğazına sahip maskelenmiş atlama bağlantıları** kullanarak doğruluk-düzenlenebilirlik dengesini mükemmelleştirir. Bu, modeli eksik bölgeler için düzenlenebilir düşük frekanslı kodlara güvenmeye zorlarken, görünür kısımları kusursuzca yeniden oluşturmak için yüksek frekanslı öznitelikleri kullanır. Sonuç olarak, önceki çalışmalardaki yaygın yapaylıkların üstesinden gelinerek pürüzsüz sınır geçişleri ve öznitelige özgü düzenlemeler için kontrol korunur. Bu katkılar bir araya geldiğinde, üretken çekişmeli ağ tabanlı görüntü içboyamayı ileriye taşıyan sağlam bir çözüm sunar. Sunulan yöntem, öznitelik uzayında çalışarak ve bilgi akışını yöneterek çeşitli, yüksek doğruluklu ve düzenlenebilir içerik üretmede üstün başarı gösterir. Kapsamlı deneyler, hem kalite hem de çeşitlilik açısından mevcut yaklaşımları geride bıraktığını ve görüntü içboyama ile düzenleme için hızlı, kontrol edilebilir bir sistem sunduğunu göstermektedir.

Anahtar sözcükler: Çekişmeli Üretken Ağlar (GAN'lar), Görüntü İçboyama, GAN Ters Çevirme, Style-GAN, Üretken Modeller, Görüntü Düzenleme, Çeşitli Üretim.

Acknowledgement

I wish to extend my sincere gratitude to my advisor, Asst. Prof. Ayşegül Dündar Boral, for her invaluable guidance, unwavering support, and mentorship throughout the course of my academic journey.

I appreciate Prof. Ugur Güdükbay and Assoc. Prof. Ramazan Gökberk Cinbiş for accepting to be my thesis jury.

I also wish to express my heartfelt appreciation to my friends Yiğit Ekin, Ahmet Burak Yıldırım, and Halil Arda Özongun, whose support, encouragement, and companionship have not only sustained me throughout this journey but have also made the process both enjoyable and memorable.

Contents

1	Introduction	1
2	Preliminaries and Related Work	3
2.1	Image Inpainting	3
2.2	Generative Models for Image Synthesis	5
2.2.1	Generative Adversarial Networks (GANs)	5
2.2.2	Diffusion Models	8
2.3	GAN Inversion and Image Editing	9
3	Method	12
3.1	Architecture	14
3.2	Architecture Details	18
3.3	Objectives and Training	20
4	Experiments	22

4.1 Experiments	22
5 Conclusion and Future Work	35



List of Figures

2.1	An illustration of the image inpainting task [1].	4
2.2	A diagram [2] illustrating the architecture of a Generative Adversarial Network (GAN).	6
2.3	The architecture of the StyleGAN generator.	7
2.4	The two processes of a Diffusion Model.	8
2.5	An illustration of GAN Inversion and subsequent latent space manipulation.	11
3.1	Comparison of the diversity behavior of various approaches. . . .	13
3.2	Feature generation initialization.	16
4.1	Qualitative results of our and competing methods on the FFHQ dataset.	22
4.2	Diversity results of various approaches.	32
4.3	Qualitative results of our diffusion-based RePaint and CoPaint models.	32

4.4	Qualitative results of AFHQ-Dog, Stanford-Car, and BreCaHAD dataset trainings.	33
4.5	Qualitative results of Ablation Study.	33
4.6	Editing results of our method.	34
4.7	The results of utilizing state-of-the-art inpainting and editing methods regarding the preservation of unerased pixels.	34
4.8	Comparison of our approach with recent state-of-the-art diffusion-based models.	34

List of Tables

4.1	Quantitative results of our and competing methods on the FFHQ validation dataset.	24
4.2	Quantitative results of StyleGAN-based methods on AFHQ-Dog and Stanford-Car datasets.	24
4.3	Runtime (in minutes:seconds:milliseconds format) and GFlops of our models compared with diffusion-based models.	25
4.4	Quantitative results of our ablation study on the FFHQ validation dataset.	33

Chapter 1

Introduction

Advances in generative modeling have opened exciting new frontiers in visual computing, particularly in image inpainting. Much of this progress is driven by powerful architectures like Latent Diffusion Models and Generative Adversarial Networks (GANs). While diffusion models have shown promising capabilities, especially when guided by text [3], GANs remain a popular choice for unconditional inpainting tasks due to their inference speed and model efficiency. A central approach in this area involves inverting images into a GAN’s latent space to enable editing and restoration [4, 5, 6, 7, 8, 9]. However, this process presents a fundamental trade-off: low-rate latent spaces offer strong editability but compromise reconstruction fidelity. In contrast, higher-rate feature spaces achieve high fidelity at the cost of editability [7, 8, 9]. Striking a perfect balance between these two goals, especially for corrupted images with missing regions, remains a significant challenge.

This thesis presents a novel framework that addresses these fundamental limitations in GAN-based image inpainting. The proposed work introduces a new architecture designed to unify high-quality, diverse inpainting with fine-grained editability. To achieve this, the proposed method leverages a dual-pathway generation process where features from an encoded image and a random sample are mixed directly within StyleGAN’s feature space (F), a key distinction from

prior work that operates in the latent space [10, 11, 12]. Furthermore, we introduce masked skip connections that create an information bottleneck, carefully balancing the need for high-fidelity reconstruction of visible areas with preserving editability and diversity for inpainted regions.

The following chapters will provide the preliminaries, a review of related work, a detailed presentation of the proposed framework, and a conclusion summarizing the contributions and findings.

Chapter 2

Preliminaries and Related Work

This chapter provides a foundational overview of the generative modeling techniques and related work that form the basis of this study. The chapter introduces two cornerstone paradigms in modern image synthesis: Generative Adversarial Networks and Diffusion Models. Subsequently, it delves into the specific application of image inpainting, discussing current state-of-the-art approaches and their limitations. Finally, the task of GAN inversion and latent space editing is framed, which is central to achieving controllable and diverse image completion.

2.1 Image Inpainting

Image inpainting is the task of filling in missing or corrupted regions of an image [1]. Given an image $\mathbf{x}$ and a binary mask $\mathbf{M}$ (where 1 indicates a missing pixel), the goal is to generate a plausible and contextually coherent completion for the masked regions. The input is the corrupted image $\mathbf{x}_{\text{masked}}$, which can be represented as:

$$\mathbf{x}_{\text{masked}} = \mathbf{x} \odot (1 - \mathbf{M}) \quad (2.1)$$

where $\odot$ denotes element-wise multiplication. The challenge of inpainting missing pixels, particularly in the case of large-scale holes, garners significant attention

from the research community and industry. This task entails the reconstruction of missing pixels in a multi-modular manner. However, many proposed methods are deterministic and struggle to address large-scale holes effectively [13, 14]. To address these challenges, CoModGAN [15] introduces an architecture and training pipeline similar to StyleGAN, incorporating skip connections from the encoder to the generator to transmit valid pixels without a bottleneck. CoModGAN achieves diverse results. Recently, the MAT model [16] trains a mask-aware transformer via adversarial learning and achieves significant results. Additionally, diffusion models have demonstrated considerable success in diverse image inpainting scenarios [17, 18, 19]. However, these models are slow due to their iterative denoising process. Furthermore, none of these methods can achieve diverse inpainting while enabling user-controlled image editing with a single network simultaneously, which constitutes the primary objective of this work.

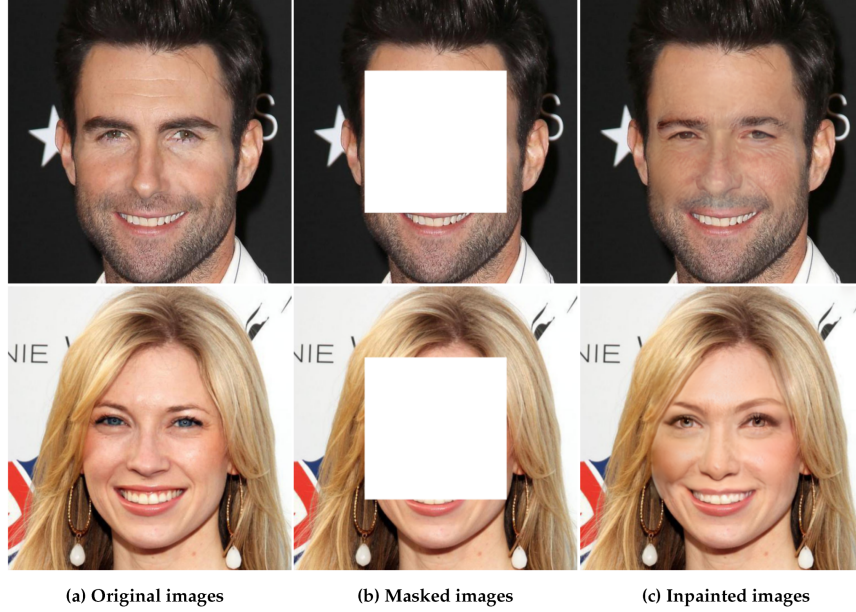


Figure 2.1: An illustration of the image inpainting task [1]. The process starts with the original images (a), which are then intentionally corrupted with a mask to create missing regions (b). An inpainting algorithm generates the final, completed images (c) by plausibly filling in the missing pixels.

2.2 Generative Models for Image Synthesis

Generative models aim to learn the underlying probability distribution of a dataset to synthesize new, similar data samples. In the context of computer vision, these models have become exceptionally powerful for tasks like image generation, translation, and editing. The focus here is on two dominant architectures: GANs and Diffusion Models.

2.2.1 Generative Adversarial Networks (GANs)

Image inpainting involves replenishing erased pixels in a manner that seamlessly integrates with the surrounding image context. While diffusion models have demonstrated promising capabilities in image inpainting, particularly when guided by text prompts to specify the object to be inserted into the erased area [3], Generative Adversarial Networks (GANs) remain a popular choice for unconditional inpainting tasks. This choice is primarily attributed to their speed and the effectiveness of their models [16].

Generative Adversarial Networks (GANs) [20] are a class of deep generative models that learn to synthesize data through a competitive process between two neural networks: a **Generator** (G) and a **Discriminator** (D). The Generator’s objective is to create data indistinguishable from real data, while the Discriminator’s goal is to identify whether a given sample is real or generated correctly.

The Generator, $G(\mathbf{z})$, takes a random noise vector $\mathbf{z}$ sampled from a prior distribution $p_z(\mathbf{z})$ (e.g., a standard normal distribution) and maps it to the data space, producing a synthetic image. The Discriminator, $D(\mathbf{x})$, takes an image $\mathbf{x}$ and outputs a scalar probability that $\mathbf{x}$ is a real image from the training dataset rather than a fake one from the Generator. This adversarial process is formulated as a two-player minimax game, described by the following objective function:

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))] \quad (2.2)$$

Here, G tries to minimize this objective while D tries to maximize it. The Generator learns to produce increasingly realistic images to "fool" the Discriminator through this adversarial training. Advanced architectures like StyleGAN [21, 22] have demonstrated a remarkable capability to organize their latent space in a semantically meaningful manner, facilitating a wide range of edits beyond what annotated datasets provide [23, 24]. This well-structured latent space is crucial for enabling intuitive and high-level image manipulation.

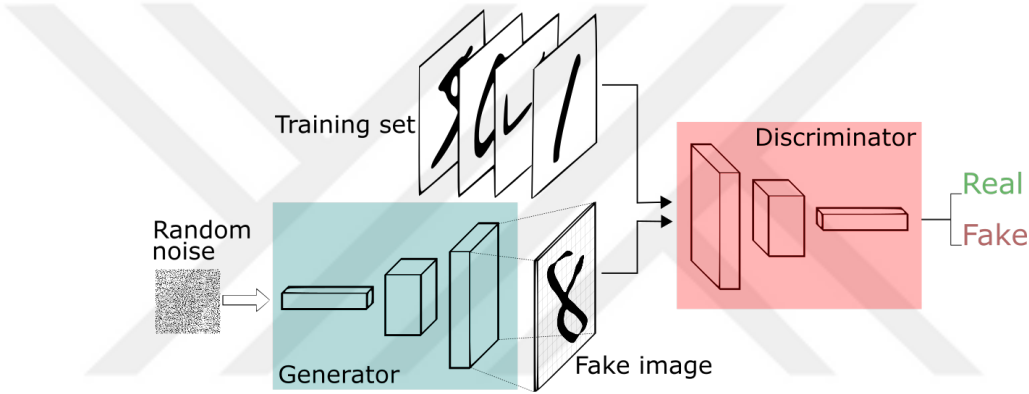


Figure 2.2: A diagram [2] illustrating the architecture of a Generative Adversarial Network (GAN). The Generator network takes a random noise vector and synthesizes an image, which is then passed to the Discriminator. The Discriminator is trained to distinguish between fake and real images from the dataset.

The StyleGAN generator architecture is designed to provide fine-grained control over the image synthesis process [21]. The process starts with a latent vector z from a standard normal distribution, which is mapped by a non-linear Mapping Network (typically an MLP) to an intermediate latent space $\mathcal{W}$ [21]. This mapping disentangles the learned factors of variation. The resulting vector w is then used to control a Synthesis Network that employs a progressive growth methodology, starting at a low resolution (e.g., 4×4) and incrementally adding layers to produce the final high-resolution image [21]. At each resolution level, w is transformed via an affine transformation into a style, which then dictates the visual features by modulating the feature maps through Adaptive Instance Normalization (AdaIN) [21]. The network's synthesis begins from a single learned constant tensor, with stochastic fine-grained details added through per-layer noise inputs. This architecture effectively separates global styles (controlled by w) from local stochastic variation (controlled by noise), enabling high-quality image generation

and intuitive editing.

GANs, particularly StyleGAN networks [22], excel in various tasks such as unconditional image generation, inpainting [12], and image editing [6, 7, 9], all within a unified framework. This versatility is made possible by the disentangled and interpretable latent space of StyleGANs. To harness the benefits of this interpretable latent space, extensive research has focused on projecting images into the GAN’s natural latent space [4, 5, 6] and exploring editing directions within the latent space [25, 24, 23]. This study introduces a framework that achieves diverse inpainting and editing using StyleGAN’s feature space. Our framework encompasses the inversion of corrupted images into the GAN’s natural latent space and the augmentation of the latent space of erased areas to enhance diversity.

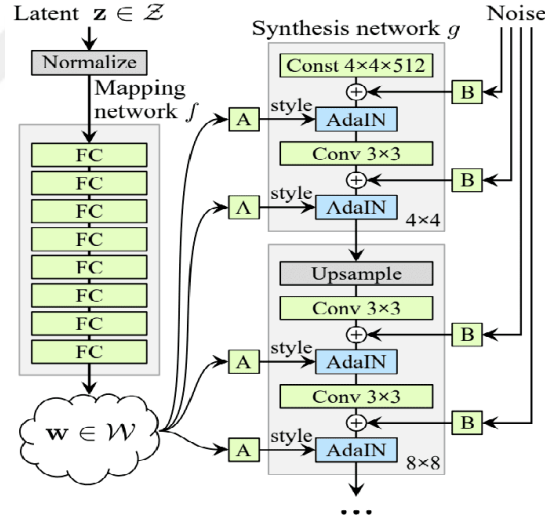


Figure 2.3: The architecture of the StyleGAN generator. A Mapping Network transforms an input latent code z into a disentangled intermediate code $w \in \mathcal{W}$. The Synthesis Network uses progressive growth, starting from a learned constant input and adding layers to increase resolution. At each level, the w vector provides style information via Adaptive Instance Normalization (AdaIN) layers, while explicit noise maps add stochastic details.

2.2.2 Diffusion Models

Diffusion Models [26, 27] represent another powerful class of generative models that learn to synthesize data by reversing a gradual noising process. The core idea involves a **forward process** that corrupts data with noise and a learned **reverse process** that restores the data from noise [28].

The forward process [28], q , systematically adds Gaussian noise to an image $\mathbf{x}_0$ over a series of T timesteps, defined as a fixed Markov chain:

$$q(\mathbf{x}_t | \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}) \quad (2.3)$$

where $\{\beta_t\}_{t=1}^T$ is a predefined variance schedule. As $t \rightarrow T$, $\mathbf{x}_T$ approaches a pure isotropic Gaussian noise distribution. The reverse process [28], p_θ , is a learned neural network trained to denoise the image by predicting the noise added at each step, starting from pure noise $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$. The training objective typically involves training a network $\epsilon_\theta(\mathbf{x}_t, t)$ to predict the added noise:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{t, \mathbf{x}_0, \epsilon} [||\epsilon - \epsilon_\theta(\sqrt{\alpha_t} \mathbf{x}_0 + \sqrt{1 - \alpha_t} \epsilon, t)||^2], \quad (2.4)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$.

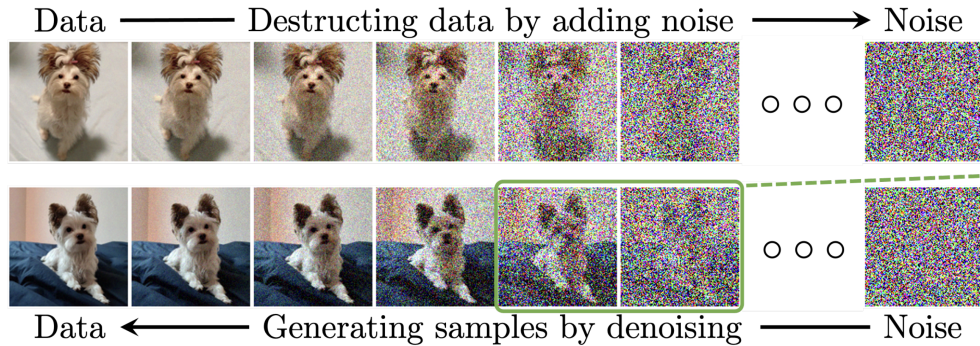


Figure 2.4: The two processes of a Diffusion Model. The forward process gradually adds noise to an image over T steps until it becomes pure noise. The reverse process learns to denoise the image, starting from noise, to generate a clean sample [28].

Diffusion-based models have emerged as a strong alternative to GANs for image inpainting tasks [18]. One advantage of GAN models is their high speed and

computational efficiency compared to diffusion models, which go through iterative denoising steps [17, 29]. Additionally, controllability remains a big challenge in diffusion models. They tend to edit the unerased pixels as well. Diffusion-based editing requires significant additional computational time for inversion [30] and does not guarantee high fidelity to the input image in every pixel. The output image may differ from the input because it passes through an auto-encoder [18], which may have difficulty perfectly reconstructing the image without an edit. Weight optimization of the decoder, achieved by training it with the reconstruction loss on a single image, is proposed for better fidelity to the input [31]. However, this approach also requires additional run-time. Moreover, the denoising steps in the encoded latent space can further alter the image [32, 33].

2.3 GAN Inversion and Image Editing

In recent times, significant attention has been devoted to image editing tasks, utilizing both end-to-end trained image translation networks [34, 35] and latent space manipulation of pretrained GANs [6, 7]. Notably, StyleGAN-based methods among pretrained GANs [22] have demonstrated a remarkable capability to organize their latent space in a semantically meaningful manner, facilitating a wide range of edits beyond what annotated datasets provide [23, 24]. However, leveraging these pretrained GANs for real image editing presents a challenge, as it necessitates the inversion of images into the natural latent space of GANs. This process, known as **GAN Inversion**, involves finding a latent code $\mathbf{w}$ that reconstructs a given real image $\mathbf{x}$ when passed through the generator G . The goal is to find an optimal latent code $\mathbf{w}^*$ by minimizing a loss function:

$$\mathbf{w}^* = \arg \min_{\mathbf{w}} (||G(\mathbf{w}) - \mathbf{x}||_2^2 + \lambda_{\text{LPIPS}} \mathcal{L}_{\text{LPIPS}}(G(\mathbf{w}), \mathbf{x})) \quad (2.5)$$

The process of inverting entire images into GANs’ latent space for editing purposes has become a focal point in GAN editing research [4, 5, 6, 7, 8, 9]. This focus is driven by difficulties in simultaneously achieving high-fidelity image reconstruction and maintaining quality in image editing. It has been observed that projecting images into low-rate latent spaces (such as z , W , or W^+ for StyleGANs) poses a

challenge in terms of faithfully reconstructing inputs due to information bottlenecks. However, these low-rate latent spaces still offer the potential for effective image editing. Conversely, higher-rate StyleGAN feature spaces allow for high-fidelity reconstruction but necessitate mechanisms to preserve the editability of low-rate codes [7, 8, 9]. While striking a balance between these factors remains challenging in the context of complete image inversion, an even more formidable task is achieving this balance for corrupted images.

Recently, several GAN inversion-based image inpainting methods have emerged, including InvertFill [10], DualPath [11], and DivInv [12]. InvertFill and DualPath utilize an image encoder trained to predict W^+ codes along with skip connections to a pretrained StyleGAN model. However, these methods approach image inpainting deterministically and cannot edit images using the encoded W^+ space. DivInv is the closest to our approach, achieving diverse inpainting and editing simultaneously. However, its design involves a mixing network that combines randomly sampled latent codes with encoded ones from images in low-rate W^+ space. We have observed that learning to mix features in W^+ space is challenging for the encoders, leading to difficulties in achieving seamless inpaintings.

Once an image is inverted to its latent code $\mathbf{w}^*$, **latent space editing** can be performed by manipulating this code along semantically meaningful directions:

$$\mathbf{w}_{\text{edit}} = \mathbf{w}^* + \alpha \mathbf{d}, \quad (2.6)$$

where $\mathbf{d}$ is a vector representing a semantic attribute (e.g., "add smile") and α is a scalar controlling the intensity. This study aims to tackle a more complex inversion problem involving input images with missing pixels, commonly called image inpainting, aiming to simultaneously inpaint and edit within a single, unified framework.

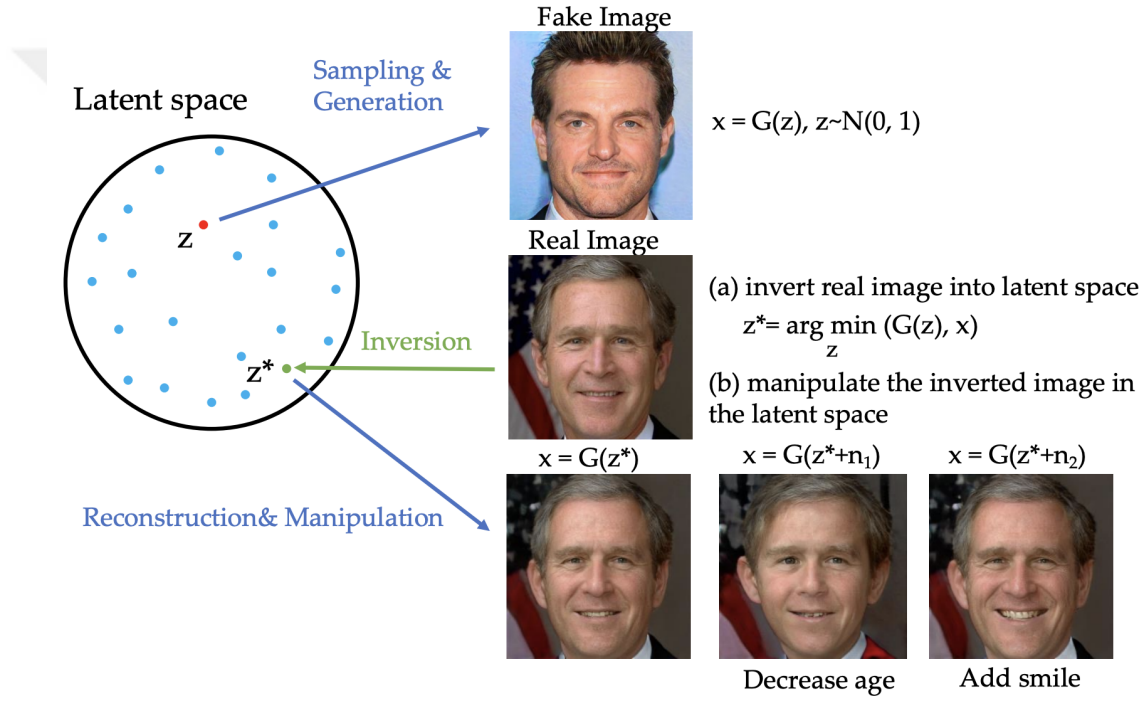


Figure 2.5: An illustration of GAN Inversion and subsequent latent space manipulation. The top path shows the standard generation process where a random latent code $z \sim \mathcal{N}(0, 1)$ is sampled to create a synthetic image. The main workflow involves GAN Inversion (a), where a real image is projected into the latent space to find an optimal code z^* that reconstructs it. Following the inversion, this code can be manipulated (b), for instance, by adding a direction vector n , to perform semantic edits such as decreasing age or adding a smile, as shown in the bottom row.

Chapter 3

Method

This chapter details the proposed methodology, beginning with an introduction to the novel architecture and an outline of the loss objectives used for training.

The proposed work introduces a novel approach involving mixing features in StyleGAN’s feature space (F), resulting in highly diverse and high-quality inpaintings. The framework is designed to preserve image editability, unifying the tasks of inpainting and editing within a single model.

The main contributions of this work are as follows:

- A novel architecture encodes a corrupted (erased) image into the W^+ space and simultaneously samples another W^+ vector from the mapping network. During the generation process, features from these two vectors are mixed at each StyleGAN layer, and a dual feedback mechanism is applied to the resulting features after each mixing step.
- To balance the trade-off between editability (from low-rate latent codes) and fidelity (from high-rate features), the architecture incorporates skip connections from the encoder to StyleGAN’s feature maps, regulated by an information bottleneck. Specifically, the erased area within the skip connections is masked, ensuring that the completion of these regions relies on the editable low-rate W^+

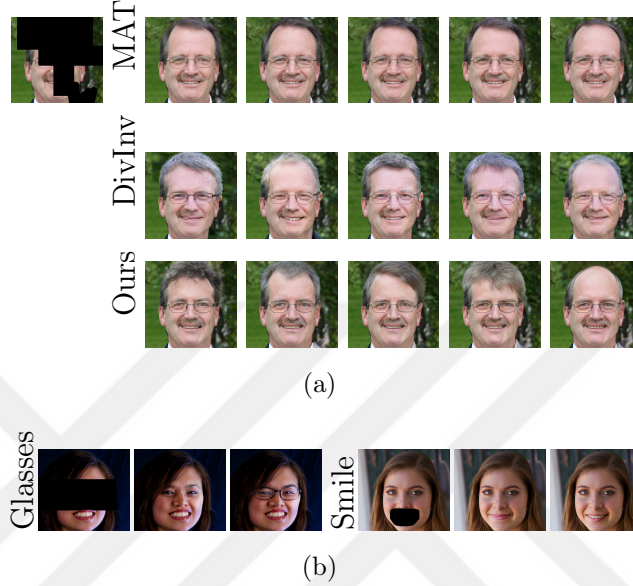


Figure 3.1: Comparison of the diversity behavior of various approaches. When a portion of the face is removed, the leading MAT model [16] exhibits a modest level of diversity. In contrast, DivInv [12] produces varied outcomes, albeit sometimes not perfectly aligned with the original content. Our approach achieves remarkable diversity, seamlessly fills in the missing areas, and offers control over the inpainted region, leveraging the interpretable latent space of StyleGAN, which is demonstrated by adding eyeglasses or smile attributes to corrupted images via erasing.

features. This mechanism preserves editability and diversity while enhancing the seamless transition between erased and unerased regions.

- Extensive experiments demonstrate the framework’s effectiveness. Notable outcomes include outstanding diversity and the seamless completion of missing areas, as illustrated in Fig. 3.1. The StyleGAN-based nature of the method ensures it is fast and allows for control over filling attributes—a combination of advantages not found simultaneously in previous methods. Furthermore, extensive comparisons are made against recent state-of-the-art diffusion models for both inpainting and editing tasks.

3.1 Architecture

Our framework includes a trainable image encoder (E), mixing networks (Mi), and frozen StyleGAN2’s mapping (Ma) and generator (G) networks [22] as shown in Fig. 3.2.

We create corrupted images by applying a binary mask M to them, where M specifies which pixels are valid and which are erased: $I^e = M \odot I$. Before passing the erased image to the encoder, we append the binary mask M . We utilize a basic encoder [36] to map our corrupted images. These steps are standard in the previous StyleGAN inversion-based image inpainting methods as well [10, 11, 12]. The rest of the framework significantly differs from the previous methods.

Our goal is to generate encoded features via powerful StyleGAN models. However, if features are only encoded from the corrupted image, that will result in a deterministic model. We establish a secondary pathway dedicated to handling sampled W^{rand} to achieve diversity. While DivInv [12] attempts to merge W^{enc} and W^{rand} into low-rate latent codes, we pursue a distinct strategy. In our approach, both W^{enc} and W^{rand} are simultaneously passed through StyleGAN to initiate feature generation. Our framework facilitates communication between these two pathways, allowing the generated features from the second pathway to augment the first pathway (illustrated in Fig. 3.2) and introduce diversity. Moreover, the second pathway needs to remain informed of the progress made by the first pathway. For instance, if the initial layers establish the pose of a person, the second pathway should align with this pose while maintaining diversity in other features. To ensure awareness of both pathways, we employ mixing networks that integrate features from both paths, with their outputs being replaced by the generated features from StyleGAN at each layer.

Our method works on unaligned datasets because the random branch aligns based on the generated ones via our novel pipeline. For instance, as illustrated in Fig. 3.2, the random branch might generate a face with a different pose, wearing eyeglasses, or even depict an older version compared to the erased input image.

At the initial stage, the second path has no knowledge of the erased image, and we present intermediate results to explore what would happen if the mixing were stopped at that point. In this case, the generations would continue solely using W^{enc} and W^{rand} , without further integration of the erased input image. If the image is largely erased, the mixing network can rely entirely on the generation from the random path. However, if only part of the face is erased, the pose information must come from the encoded branch, which feeds back into the random path. The random path, in turn, adapts by taking the pose from the encoded features but continues with its own random attributes, such as eyeglasses or other personal traits. The mixing network can integrate the erased features from this random path based on the erasing mask and encoded features. As the generation progresses through the later layers of StyleGAN, the outputs of both paths begin to converge, since, after mixing, they continue using the same features.

Specifically, at each layer, we take the features from StyleGAN for F^{rand} , F^{enc} , and the mask M and feed them to a small encoder-decoder-based network. F^{rand} and F^{enc} correspond to output features of convolution layers in StyleGAN at different resolutions. The mask is downsampled to match the features' resolution.

$$F^{comb} = \text{NN}(F^{enc}, F^{rand}, M) \quad (3.1)$$

The next layer features are calculated after replacing both F^{rand} and F^{enc} with F^{comb} and continue the convolutional and adaptive instance normalization layers with W^{rand} and W^{enc} , respectively. This way, F^{comb} can decide on the person's pose based on what is visible and what can be taken for diversity. Once the pose is established, the second pathway must generate aligned features for the given pose to provide diversity for attributes like facial features and hair colors. This is the reason we also use F^{comb} in the second path, but keep generating new features with W^{rand} . We observe that this method achieves better alignment with the unerased pixels' content than DivInv while achieving diversity.

The second component of our framework to achieve editability, diversity, and

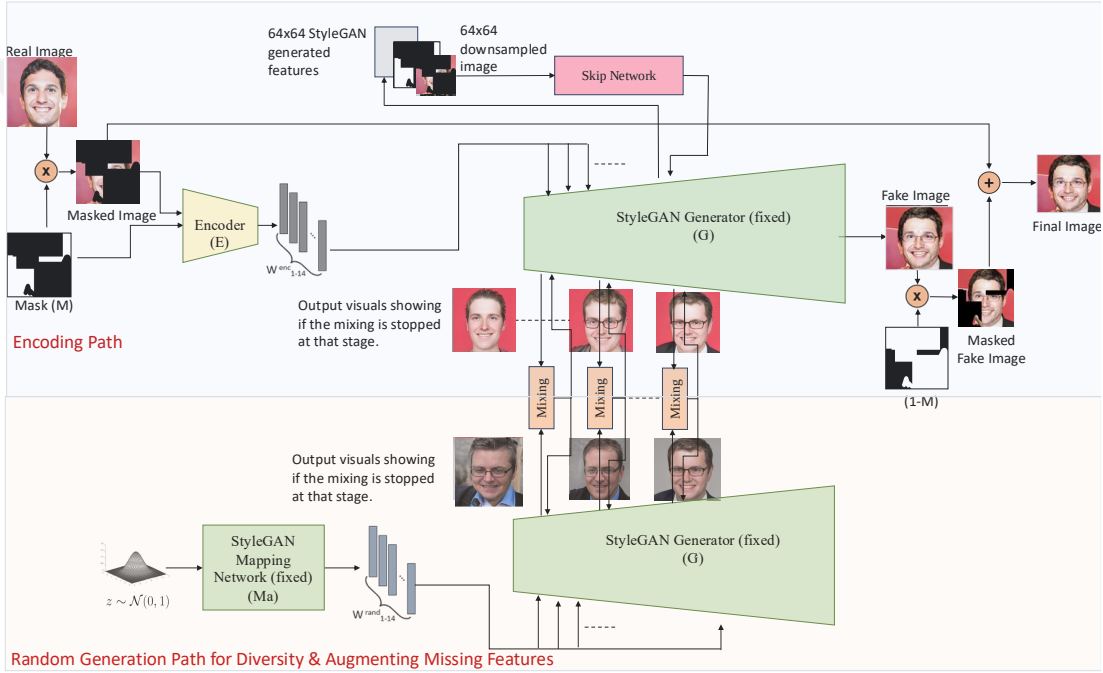


Figure 3.2: Feature generation initialization. Our approach involves passing both W^{enc} and W^{rand} through StyleGAN simultaneously to initiate feature generation. Mixing networks are employed to integrate features from both paths, with their outputs replaced by generated features from StyleGAN at each layer to ensure awareness of both pathways. This approach allows communication between the two pathways, enabling features generated by the second pathway to enhance the first pathway and introduce diversity. Additionally, the second pathway must remain informed of the progress of the first pathway, ensuring alignment with established features while maintaining diversity.

seamless inpainting is the skip connections from high-rate encoder features to the StyleGAN generator with an information bottleneck. Previous research on image inversion has highlighted the limitations of low-rate latent codes, such as W^+ , in achieving high-fidelity image reconstruction [7, 8]. The bottleneck of projecting high-dimensional images onto these low-rate latent codes results in significant information loss, constraining the potential for high-fidelity reconstructions. This limitation is also evident in image inpainting tasks, as observed in previous works [11, 10]. In inpainting, the accurate reconstruction of visible pixels is particularly crucial because the final output combines erased and generated pixels via a mask, as shown in Fig. 3.2 (final image generation from fake image). When the generator fails to reconstruct valid high-fidelity pixels, noticeable color discrepancies arise in the final images.

Earlier unerased image inversion techniques [7, 9] propose encoding information into higher-rate latent codes. This task is accomplished through skip connections from the encoder to the generator, operating on higher-resolution feature maps known as F space. The optimal distribution of information between W^+ and F spaces is a subject extensively explored in image inversion approaches [7, 8, 9]. Emphasizing F space for encoding ensures high-fidelity reconstruction, yet it may compromise editability as the image is not appropriately projected into W^+ . Achieving a balance is crucial, especially for inversion methods for effective image editing. All these image inversion methods [7, 8, 9] first train an encoder to learn W^+ and then freeze it to learn skip connections, ensuring the encoder learns to project images to W^+ and does not cheat by passing information via F. In our case, we expect editability for the erased parts; for example, when lips are erased, we aim to generate diverse lips and manipulate them to smile by user control. Therefore, we intend to heavily rely on inpainting the erased parts using features from W^+ . On the other hand, F space features are crucial for our setting because, in the end, we stitch the unerased image with the inpainted one. Thus, slight color differences or discontinuities in the hair become even more visible than in the unerased image inversion task.

To strike this balance, we introduce skip connections from the encoder to the generator in the 64×64 dimensional F space. However, we mask these skip

connections with the input mask downsampled to the same resolution as the features. Without this masking, the framework does not incentivize learning the W^+ features, relying solely on the encoder-decoder setting to fill the inpainted regions. Conversely, if these connections are entirely removed, the model struggles to achieve seamless boundaries when inpainting. Therefore, we find that skip connections with a bottleneck are a critical component for the success of our method. We feed these skip connections at 64×64 resolution to the mixing network operating on that layer. The mixing network at that layer takes skip features from the encoder in addition to the F^{rand} and F^{enc} . Additionally, we incorporate a convolutional second-stage network into our training process. This network is designed to process the erased parts further, aiming to achieve smoother transitions between erased and unerased pixels, similar to the approach used in MAT and DivInv models.

The final images are obtained as follows:

$$I^f = M \odot I + (1-M) \odot I^o \quad (3.2)$$

where I^o is the output of the network. This step guarantees that the valid pixels are not modified, which is standard practice in inpainting tasks.

3.2 Architecture Details

In this section, we present the architecture details.

Encoder (E). We adopt the encoder architecture from pSp with minor modifications. First, we increase the first layer input channel number from 3 to 4 to take the mask as an additional input. Then, we disable the normalization layers since we observe they decrease the model’s performance, given that many input pixels may be 0 due to their removal.

Feature Space Mixing Network (*Mixing*). We employ mixing networks that operate on 8×8 , 16×16 , 32×32 feature maps of the StyleGAN2 generator. First, two 8×8 feature maps are generated independently, by W^{enc} and W^{rand} . At this point, we employ mixing networks with the U-Net architecture. The 8×8 feature maps have 512 channels; we concatenate and feed them into the mixing network. The mixing network produces the 512-channel combined feature map output with a spatial resolution of 8×8 . StyleGAN2 generator generates the next feature map (with spatial resolution of 16×16) generated by conditioning on this 8×8 combined feature map. Rather than creating independent feature maps for W^{enc} and W^{rand} , we use the combined map and inject styles of W^{enc} and W^{rand} . This procedure enables the network to both preserve the diversity and obtain a better merging between feature maps of W^{enc} and W^{rand} . The same mixing procedure goes up to (including) resolution 32×32 . The U-Net architecture has three downsampling and three upsampling blocks, where downsampling blocks reduce the resolution by half and upsampling doubles it. U-nets first reduce the number of channels to 64, then at each downsampling block, double the number of channels and halve the number of channels at each upsampling block.

Skip Network. We employ a Skip Network with the same U-Net architecture used in Mixing networks described above. The Skip Network only operates on 64×64 spatial resolution. We downsample the masked input image and the mask to 64×64 resolution, and concatenate them to the 64×64 feature map generated by StyleGAN2’s generator. We feed the 516 (512 for feature maps, 3 for masked input, 1 for mask) channel input to the U-Net, and the U-Net outputs the 512 channel output. At this resolution, no mixing network is applied.

Second Stage Network. For the second stage, we use a U-Net model, which downsamples images to 16×16 resolution and upsamples back to 256×256 . This stage improves the color match between the inpainted and original pixels. The number of channels in the first layer output is 128, increasing 2 times at each layer up to 512. In the upsampling layers, the number of channels decreases symmetrically.

3.3 Objectives and Training

We train our encoder and mixing networks using a combination of image reconstruction and adversarial losses. Following the training protocol outlined in [12], we generate synthetic images for training purposes. Initially, we generate an image with z_g and then pass it through the mapping network to obtain W_g^{rand} . Subsequently, we synthesize the output image and erase a portion of it using a randomly generated mask, denoted as $I_g^e = M \odot I_g$. Our encoder then processes the corrupted (erased) image to obtain W_g^{enc} . The generator produces images I_g^o and I_r^o via these using the same W_g^{rand} that generated the image and a new randomly sampled one. The generation of I_g^o benefits from complete information that generated the image, allowing for accurate reconstruction when combined with the encoded embedding. In contrast, I_r^o only has access to information from unerased pixels and is not expected to reconstruct the input image for the erased regions faithfully. To achieve these objectives, we set the following losses.

Reconstruction Losses. For reconstructing the pixels correctly, we use L_2 and perceptual losses (VGG (Φ)). We expect I_g^o to reconstruct I_g but I_r^o to only match I_g for the unerased pixels. Therefore, when we calculate reconstruction losses for I_r^o , we mask out the erased pixels. On the other hand, for I_g , there is no masking out any pixels.

$$\begin{aligned}\mathcal{L}_{rg} &= \|I_g^o - I_g\|_2 + |\Phi_j(I_g^o) - \Phi_j(I_g)|_2 \\ \mathcal{L}_{rr} &= \|(M \odot I_r^o) - I_g^e\|_2 + |\Phi_j(M \odot I_r^o) - \Phi_j(I_g^e)|_2\end{aligned}\tag{3.3}$$

Adversarial Losses. We additionally guide the training with adversarial loss to achieve realism. As in the StyleGAN inversion methods, we load the pretrained discriminator from StyleGAN training. For real images, the discriminator takes the original dataset images, and for fake images, we feed the final outputs obtained as given in Fig. 3.2.

Mean W Loss. We do not want encoded W^+ deviate from the natural

StyleGANs latent space. This is important to preserve the editability of the images. We average 10000 sampled W values to achieve that and apply an L2 distance loss between that and the encoded W^+ . This loss has been used by previous works as well [36].

We set the weights of reconstruction and adversarial losses to 1.0 and mean w loss to 0.01. We train our model with a batch size of 32 for 150 epochs.



Chapter 4

Experiments

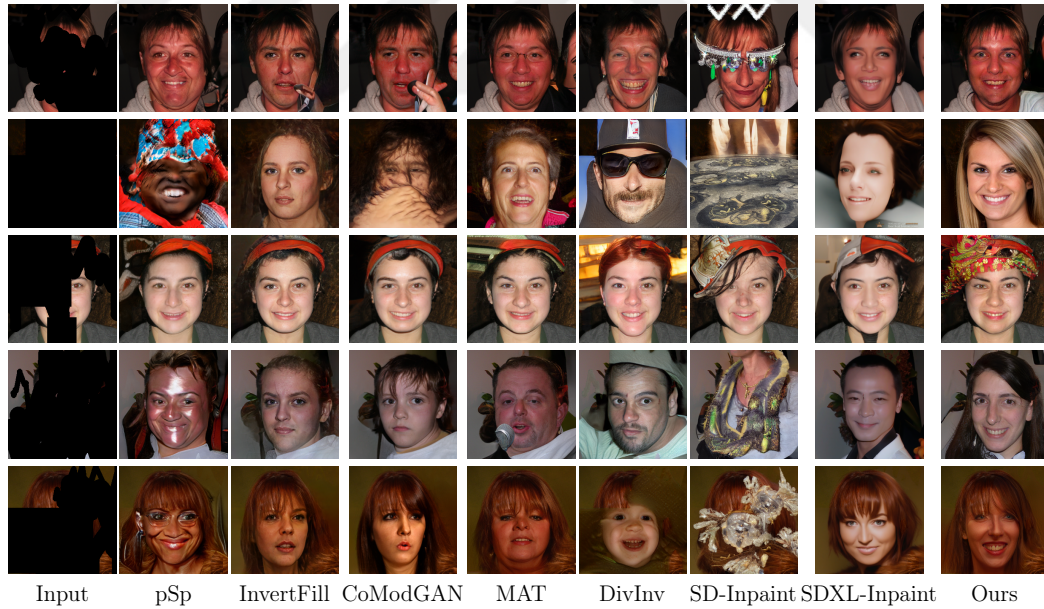


Figure 4.1: Qualitative results of our and competing methods on the FFHQ dataset.

4.1 Experiments

Baselines. We benchmark our method against leading image inversion techniques such as pSp [36], e4e [6], HFGI [7], ReStyle [37], and HyperStyle [8] that are

trained for image inpainting. We train those methods for the inpainting task using the author’s released codes. That is, we add another channel to the input network that will take the mask input, and instead of complete images, we provide masked images to the networks. Following, we evaluate InvertFill [10] and DualPath [11], which are additional StyleGAN-based image inpainting techniques. Notably, these methods operate deterministically, thus lacking diversity in their content filling. We compare with state-of-the-art image inpainting methods that achieve diversity, such as CoModGAN [15], MAT [16], and DivInv [12]. CoModGAN introduces a StyleGAN-like model, trained from scratch, incorporating co-modulation and skip connections for inpainting. MAT utilizes a two-stage network, employing mask-aware transformers in the first stage and StyleGAN-like normalization in the second stage to achieve diversity. Although these methods achieve diversity, they cannot offer editability and control over attributes during inpainting. We compare our approach with DivInv, which is the closest to our work in terms of its ability to output diverse and controllable inpaintings. However, their mixing networks operate on low-rate W^+ and lack skip connections during the first stage of training, resulting in misalignment with the unerased content.

We also compare our approach with state-of-the-art diffusion-based inpainting methods, SD-Inpaint and SD-XL-Inpaint, which are trained on millions of image-text pairs. We obtain the results for SD-Inpaint and SDXL-Inpaint models with the prompt "Realistic human face image, HD". It is important to note that this is not a direct one-to-one comparison because these methods are trained on different datasets. The SD-Inpaint methods benefit from being trained on much larger and more diverse datasets, whereas the others are trained on category-specific datasets. Nonetheless, we report these results as they are of significant interest to the community.

Lastly, we qualitatively compare our method with recent fix-diffusion-based inpainting methods, specifically CoPaint [29] and RePaint [17]. These methods, trained on the CelebA-HQ dataset, are also category-specific. We limit our comparison to qualitative analysis due to the slower speed of the compared methods, which is discussed in more detail in the Qualitative Results section.

Table 4.1: Quantitative results of our and competing methods on the FFHQ validation dataset.

Models	Div.	Edit	FID	LPIPS
e4e [6]	✗	✗	11.41	-
pSp [36]	✗	✗	8.23	-
ReStyle [37]	✗	✗	8.07	-
HFGI [7]	✗	✗	7.66	-
HyperStyle [8]	✗	✗	7.46	-
InvertFill [10]	✗	✗	7.45	-
DualPath [11]	✗	✗	17.60	-
SD-Inpaint [18]	✓	✗	29.93	0.292
SD-XL-Inpaint	✓	✗	10.05	0.145
CoModGAN [15]	✓	✗	7.35	0.079
MAT [16]	✓	✗	3.30	0.065
DivInv [12]	✓	✓	5.92	0.202
Ours	✓	✓	3.68	0.170

Table 4.2: Quantitative results of StyleGAN-based methods on AFHQ-Dog and Stanford-Car datasets.

	AFHQ-Dog		Stanford-Car		BreCaHad	
Model	FID	LPIPS	FID	LPIPS	FID	LPIPS
pSp [36]	26.46	-	11.77	-	44.35	-
HFGI [7]	20.58	-	8.55	-	31.19	-
HyperStyle [8]	23.92	-	10.97	-	37.43	-
DivInv [12]	18.89	0.173	8.39	0.193	24.20	0.401
Ours	12.07	0.292	7.53	0.242	19.43	0.406

Evaluation. Our evaluation primarily relies on two metrics: the Frechet Inception Distance (FID) metric [38] and the Learned Perceptual Image Patch Similarity (LPIPS) metric [39]. The FID metric assesses realism by comparing the distribution of target images with reconstructed images, while LPIPS measures diversity. We generate two images per input image-mask pair for models producing diverse images and calculate the distance between them using the LPIPS metric. Our goal is to achieve low FID and high LPIPS scores. We use the masks provided by DivInv authors [12] for our evaluations.

Masks range between 0 and 1, meaning the mask’s ratio can go from 0 – 100%

Table 4.3: Runtime (in minutes:seconds:milliseconds format) and GFlops of our models compared with diffusion-based models.

Models	RunTime	GFlops $\times 10^3$
RePaint [17]	02:29:000	5340
CoPaint [29]	03:21:000	2840
CoPaint-TT [29]	07:09:000	6159
SD-Inpaint [18]	00:04:000	84.17
SD-XL-Inpaint	00:04:000	2460
Ours	00:00:065	0.170

of the image. When the erasing masks are very large, the utility of using an inpainting method in such cases can be questioned. However, it is important to note that large masks are increasingly common in recent research works [16, 12], demonstrating the capability of models to handle a range of mask sizes, from smaller to larger ones. This approach showcases the robustness and versatility of the models under consideration, a significant aspect of academic exploration in this field.

Datasets. We utilize the FFHQ [22], AFHQ-Dog [40], Stanford Cars [41], and BreCaHAD [42] image datasets. Our comparisons are based on the StyleGAN2 pre-trained models [22], and we adhere to their train and validation splits. In our comprehensive comparison with existing works, we focus on the FFHQ human face dataset, which is consistent with prior research. As for the other datasets, we train the best single-stage (pSp) and two-stage encoders (HFGL, HyperStyle) for the comparisons. We also compare with DivInv, given these datasets’ lack of pretrained model availability.

Quantitative Results. We present quantitative results for FFHQ evaluations in Table 4.1, utilizing evaluation masks provided by DivInv and using their pretrained models. In this comparison, we evaluate state-of-the-art image inversion methods trained for the inpainting task, namely e4e, pSp, ReStyle, and HFGL. Although they achieve reasonable inpainting results, they lack diversity. Similarly, models designed for inpainting with StyleGAN inversion, such as InvertFill and DualPath, fail to provide diversity and editability. Next, we compare with state-of-the-art inpainting methods that offer diverse outputs, namely CoModGAN and

MAT models. These models train end-to-end networks from scratch and do not utilize StyleGAN space, thus lacking editability. While they achieve outstanding results, especially the MAT model, their diversity is limited. In our experiments, achieving diversity is challenging, as the model tends to encode everything from corrupted images rather than using sampled signals. While CoModGAN and MAT models exhibit some diversity, the changes are only minor, as measured in the LPIPS score (0.06 – 0.07).

Finally, we compare our method with DivInv, which offers diversity and editability through its StyleGAN-inversion-based inpainting network. Our method demonstrates significant improvements in FID over DivInv, as illustrated in Table 4.1. Although DivInv achieves a better diversity score, this often leads to color and structural inconsistencies with the unerased pixels. Note that high LPIPS (diversity) alone is not a meaningful metric. A high LPIPS score could indicate the generator produces images randomly, lacking alignment with the unerased content. Such a model, while diverse, would be unusable. A low FID should accompany diverse LPIPS scores.

In summary, our method matches the MAT model in FID score and outperforms it significantly in terms of diversity and editing capabilities. While maintaining similar quality, our approach opens up new possibilities by expanding the range of diversity and providing enhanced control over image manipulation for the erased areas. Additionally, our method surpasses DivInv in FID scores, establishing itself as a remarkable advancement in image inpainting and editing.

We compare with DivInv and other StyleGAN inversion-based inpainting methods on the AFHQ-Dog, Stanford-car, and BreCaHAD breast cancer histopathology images, as illustrated in Table 4.2. Our approach consistently outperforms, achieving superior results.

Qualitative Results. Figure 4.1 presents the outcomes of our method and selected competing approaches on the FFHQ dataset. Due to space limitations, we include a subset of state-of-the-art models with various approaches, such as pSp, InvertFill, CoModGAN, MAT, DivInv, SD-Inpaint, and SDXL-Inpaint.

Notably, in scenarios where the masks are extensive, as demonstrated in the second example, most methods struggle to deliver satisfactory results. SD-Inpaint achieves significantly worse results than SDXL-Inpaint. However, even the SDXL-Inpaint model struggles to generate realistic face images in those examples. InvertFill and MAT manage to produce reasonable outputs, while our method excels in generating higher-quality results. In the third example, pSp, InvertFill, CoModGAN, and MAT exhibit similar inpainting patterns, primarily extending the headband without achieving a realistic appearance. The outputs of both SD-Inpaint and SDXL-Inpaint show inconsistent pixels, with the headband filling those regions with grey accessories. Conversely, DivInv presents diverse results, yet fails to maintain seamless boundaries across the hair accessory. In contrast, our model produces elaborate and realistic hair accessories, showcasing its superior performance in this task.

We show additional diversity results in Fig. 4.2. For the same input, we inpaint the images with different random seeds. It can be seen that DivInv struggles to achieve seamless inpaintings. On the other hand, SD-Inpaint and SDXL-Inpaint do not produce realistic faces; rather, they are more like cartoon images. Our method achieves diverse inpainting results.

In our qualitative comparisons, we also include the diffusion-based models, namely RePaint [17] and CoPaint [29] models. We generate only 10 inpaintings with these methods, as each generation takes a very long time, as given in Table 4.3. The computation times are measured on Tesla T4 by averaging inference times of 10 samples. RePaint framework employs a diffusion model trained on face images and achieves diverse inpaintings by conditioning on the masked input through an iterative denoising process. CoPaint further improves the results by jointly modifying revealed and unrevealed parts of intermediate variables in each time step along the denoising process. Their additional time travel (TT) idea has improved their results. Due to the model’s lengthy runtime, we could not generate enough inpaintings to compute FID scores. However, as depicted in Fig. 4.3, these models encounter challenges in generating realistic content for many examples.

Diffusion models have become strong competitors in generative tasks. However, as highlighted, one of the main challenges they face is their high computational resource requirements, which can be prohibitive for many practical applications. In response to these challenges, recent work on single-step diffusion models has aimed to reduce computational burdens. Unfortunately, these methods still struggle to preserve the high-quality details achieved by multi-step diffusion models. Additionally, the quality of diffusion models’ results is often not on par with MixGAN’s, particularly in tasks such as image inpainting for domain-specific tasks like face, car, or biomedical image inpainting, where MixGAN outperforms diffusion models. While diffusion models trained on large-scale datasets containing millions of images can handle various image domains simultaneously, they tend to perform worse when focusing on a single domain, as in our task. In contrast, GAN-based models like MixGAN are specifically optimized for these tasks, often yielding better results in these more constrained settings. Furthermore, diffusion models generally struggle to maintain input fidelity, which can lead to border artifacts during inpainting tasks. These issues are also present in editing pipelines.

We additionally showcase the inpainting outcomes of our method on AFHQ dog images, Stanford car, and BreCaHAD breast cancer histopathology images in Fig. 4.4. Our method effectively achieves inpaints on these image domains as well.

Ablation Study. We present the quantitative results of our ablation study in Table 4.4. Throughout the development of our model, we track both diversity (LPIPS) and quality (FID) of our results. It is crucial to emphasize that high LPIPS (diversity) alone is not a meaningful metric. A high LPIPS score could indicate the generator produces images randomly, lacking alignment with the unerasd content. Such a model, while diverse, would be unusable. On the contrary, FID scores reflect the quality of the inpainted images, requiring semantic consistency and seamless boundary transitions. During our development, we strive to enhance FID while maintaining diversity.

We use DivInv as our base model for the ablation study. They propose improvements to the mixings in W^+ latent codes. They also employ a two-stage network, common with GAN-based image inpainting tasks. A similar refinement

mechanism is used by the state-of-the-art MAT model as well. We compare our first-stage model with DivInv. By replacing their mixing mechanism with our feature space-based mixing, we significantly improved FIDs, 16.65 versus 10.33, while achieving similar diversity. By incorporating additional masked skip connections, the model demonstrates improved alignment with the content, as evidenced by the FIDs. The masked connections further enhance diversity scores, likely due to providing the network with clearer information about visible features, facilitating better integration with diverse features. Finally, adding a second stage to our model brings our FIDs to 3.68, which is significantly better than 5.92, the score of the final DivInv model. We also present an ablation study in which we remove one of the loss components described in Section 3.2. Omitting any of the losses leads to significant degradation in performance.

Do we need a feature mixing mechanism? While StyleGAN’s style-mixing functionality is well-established, it falls short in addressing diversity in inpainting tasks where consistency with the unerased images is a must. Style-mixing involves obtaining different z samples and generating a mixed W^+ code. To illustrate, we have included outputs in Fig. 4.5. As seen in these examples, random mixing leads to inconsistent inpainting results. For instance, encoding W^+ vectors from corrupted images and replacing the W^+ vectors with randomly sampled ones in the later layers results in color inconsistencies, as evident in the first example. Similarly, using earlier layers produces erroneous content, such as the infilled area appearing younger than the realistic output, as seen in the second example. Next, we also show that if we remove the feedback to the second pathway and let the generator generate a random image, the model does not learn to use the second pathway because it cannot align them. It falls to using only the encoded path, hence outputs deterministic results as shown in Fig. 4.5 (w/o feedback mechanism results). Considering these points, removing the mixing mechanism in feature space or removing the feedback to the random path eliminates the diversity, underscoring our architecture’s significance. Finally, we present visual results from our experiments, comparing the performance with and without masked skip connections. These results highlight that the method fails to properly fill the space aligned with the unerased parts without the skip connection.

Editing Results. In our editing process, we begin by encoding the corrupted image and then randomly sample a W^+ vector to diversify the generation process. Subsequently, we edit the W^+ vectors using InterfaceGAN [25] and StyleClip [23] directions. The editing results are presented in Fig. 4.6. Initially, we display the corrupted (erased) input. Following this, the subsequent rows depict the results before and after applying editing directions for various attributes. Our method leverages the latent space of StyleGAN to enable controllable editing. Extensive research into StyleGAN’s latent space has uncovered numerous editable directions, which are directly applicable to our method, as we utilize a pretrained StyleGAN generator.

Do we need a unified framework for inpainting and editing tasks? Unified frameworks offer efficiency and robustness and have garnered significant interest in machine learning research. However, from an application standpoint, one might question the advantage of achieving both inpainting and editing under a single framework. Specifically, the question arises: can we separately inpaint images and then edit them using StyleGAN inversion via state-of-the-art networks? However, using that approach not only utilizes two networks, each of which is the same size as ours, but also overlooks one of the key promises of inpainting: preserving unerased pixels. In inpainting tasks, unerased pixels are seamlessly integrated into the final image via a masking operation as given in Eq. 3.2. However, as illustrated in Fig. 4.7, editing methods do not maintain this guarantee. Upon incorporating the original pixels post-editing, noticeable boundaries emerge, detracting from the visual coherence of the image. Therefore, the unified framework offers efficiency and opens up new possibilities for applications.

With advancements in diffusion-based editing techniques, are StyleGAN-based editing frameworks still relevant? In this study, we introduce a framework capable of achieving inpainting and editing in a single feed-forward process. We demonstrate that while stable diffusion models show promising results, StyleGAN models trained on specific categories offer advantages not only in speed but also in the quality of inpainting results. Next, we investigate the editing capabilities of diffusion versus StyleGAN models. We compare our approach with recent state-of-the-art diffusion-based models, namely LEDITS++

[33], NoiseCLR [32], and InfEdit [43] in Fig. 4.8. We inpaint and subsequently edit images using our method and others. We avoid masking operations for the other methods to blend original unmasked pixels with the generated ones. This omission is deliberate as diffusion-based inpainting methods alter all pixels, often lacking fidelity to the input and struggling with edits. If we add the masking operation as given in Eq. 3.2, the results would become worse.





Figure 4.2: Diversity results of various approaches. Our approach stands out as the sole method capable of achieving remarkable diversity while seamlessly filling in missing areas simultaneously.

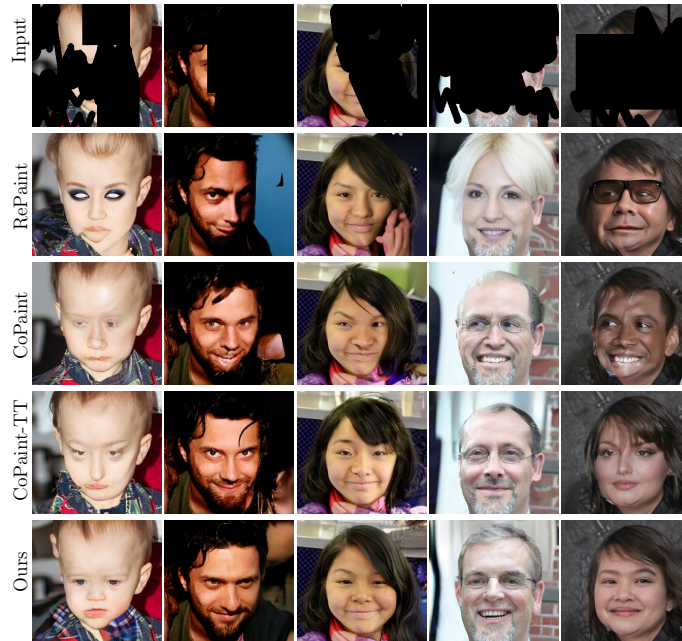


Figure 4.3: Qualitative results of our diffusion-based RePaint and CoPaint models.



Figure 4.4: Qualitative results of AFHQ-Dog, Stanford-Car, and BreCaHAD dataset trainings.

Table 4.4: Quantitative results of our ablation study on the FFHQ validation dataset.

Methods	Second Stage	FID	LPIPS
DivInv		16.65	0.1440
DivInv	✓	5.92	0.2026
Feature Space Mixing (FSM)		10.33	0.1243
FSM + Masked Skip connection		7.99	0.1723
FSM + Masked Skip connection	✓	3.68	0.1692
w./o. adversarial Loss	✓	12.35	0.1704
w./o. reconstruction Loss	✓	16.02	0.1668
w./o. meanw Loss	✓	13.80	0.2070

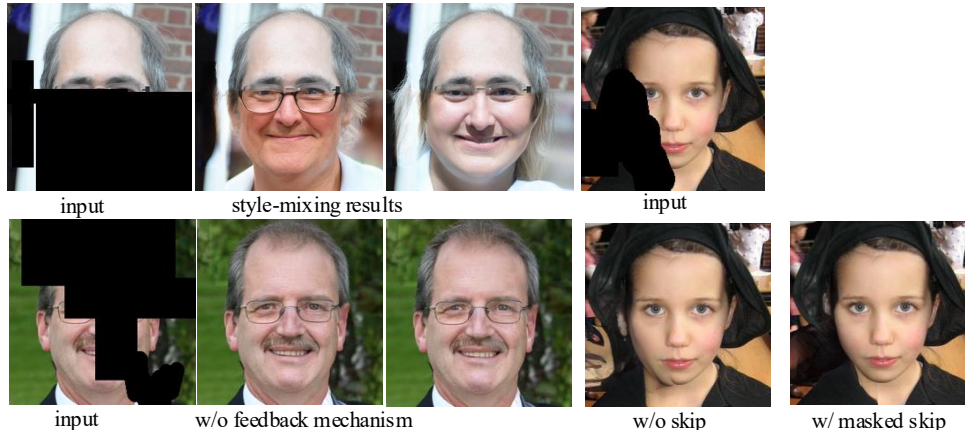


Figure 4.5: Qualitative results of Ablation Study.



Figure 4.6: Editing results of our method. The second row shows the randomly sampled W+ results, and the third row shows the results after adding InterfaceGAN (I) and StyleClip (S) directions for different attributes.

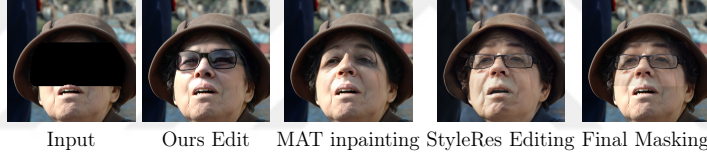


Figure 4.7: The results of utilizing state-of-the-art inpainting and editing methods regarding the preservation of unerased pixels. If we opt for utilizing state-of-the-art inpainting and editing (StyleRes [9]) techniques instead of our integrated approach, achieving both tasks, it is important to note that typical editing methods do not assure the preservation of unerased pixels. Consequently, reintroducing the original unerased pixels into the output as given in Eq. 3.2 may lead to boundary artifacts.



Figure 4.8: Comparison of our approach with recent state-of-the-art diffusion-based models. We inpaint and subsequently edit images using our method and others. We avoid masking operations to other methods to blend original unmasked pixels with the generated ones (as given in Eq. 3.2). This omission is deliberate as diffusion-based inpainting methods alter all pixels, often lacking fidelity to the input and struggling with edits.

Chapter 5

Conclusion and Future Work

This study introduced a novel framework designed to address a central challenge in image inpainting: the simultaneous achievement of high-quality, diverse, and editable results. By harnessing the powerful and interpretable feature space of a pretrained StyleGAN2 generator [22], the proposed method successfully unifies these three often-conflicting objectives within a single, efficient feed-forward network. The core of this approach lies in a dual-pathway architecture that intelligently mixes features from an encoded input and a random latent vector within StyleGAN’s high-dimensional F space. This feature-space mixing—a departure from prior works that operate in the low-rate W^+ space [12]—augmented by masked skip connections and a two-stage refinement process, was shown to be critical for success, as validated by a thorough ablation study where each component demonstrably improved performance.

The framework’s effectiveness was rigorously evaluated through extensive quantitative and qualitative experiments across multiple datasets. The results confirm that the method sets a new standard for controllable GAN-based inpainting. Quantitatively, the approach achieves an FID score of 3.68 on FFHQ, highly competitive with the state-of-the-art deterministic model MAT [16] (3.30) while offering the crucial, previously missing capabilities of diversity and editability.

Regarding generation quality and realism, it significantly outperforms the closest comparable method, DivInv [12] (5.92 FID). Furthermore, the framework demonstrates robust generalization, consistently achieving superior FID scores on non-face datasets like AFHQ-Dog and Stanford Cars.

Qualitative comparisons highlight the method’s ability to generate seamless, coherent, and highly diverse completions. Other methods, including leading diffusion models like SD-XL [18], often fail to produce realistic or domain-consistent results. A key advantage of this GAN-based approach is its exceptional computational efficiency, with an inference time of just 65 milliseconds, orders of magnitude faster than diffusion-based competitors [17, 29], which can take several minutes. This makes the framework practical for real-world applications. The integrated editing capabilities, leveraging the rich semantics of StyleGAN’s latent space with established editing directions [25, 23], were shown to be superior to a sequential inpaint-then-edit pipeline, which often introduces visual artifacts at object boundaries.

While the proposed framework demonstrates significant advantages, its primary limitation is its reliance on a domain-specific, pretrained GAN, which confines its application to the data distribution it was trained on (e.g., faces, cars). Future work could explore hybrid approaches integrating insights from large-scale diffusion models to broaden the framework’s domain adaptability without sacrificing the fidelity and efficiency demonstrated here. Further research could also focus on developing more fine-grained editing controls directly within the feature-mixing process.

In conclusion, this work successfully bridges the gap between high-fidelity reconstruction, generative diversity, and semantic control in image inpainting. Establishing a new benchmark in performance and capability provides a comprehensive and efficient solution that opens up new possibilities for creative and practical image manipulation tasks.

Bibliography

- [1] L. He, Z. Qiang, X. Shao, H. Lin, M. Wang, and F. Dai, “Research on high-resolution face image inpainting method based on StyleGAN,” *Electronics*, vol. 11, no. 10, p. 1620, 2022. [viii](#), [3](#), [4](#)
- [2] Thalles, “A short introduction to generative adversarial networks.” <https://sthalles.github.io/intro-to-gans/>, 2017. Accessed: 2024-08-08. [viii](#), [6](#)
- [3] O. Avrahami, D. Lischinski, and O. Fried, “Blended diffusion for text-driven editing of natural images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18208–18218, 2022. [1](#), [5](#)
- [4] A. Creswell and A. A. Bharath, “Inverting the generator of a generative adversarial network,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 7, pp. 1967–1974, 2018. [1](#), [7](#), [9](#)
- [5] 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. [1](#), [7](#), [9](#)
- [6] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, “Designing an encoder for StyleGAN image manipulation,” *ACM Transactions on Graphics*, vol. 40, no. 4, pp. 1–14, 2021. [1](#), [7](#), [9](#), [22](#), [24](#)
- [7] T. Wang, Y. Zhang, Y. Fan, J. Wang, and Q. Chen, “High-fidelity gan inversion for image attribute editing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11379–11388, 2022. [1](#), [7](#), [9](#), [10](#), [17](#), [22](#), [24](#)

- [8] Y. Alaluf, O. Tov, R. Mokady, R. Gal, and A. Bermano, “HyperStyle: StyleGAN inversion with hypernetworks for real image editing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18511–18521, 2022. [1](#), [9](#), [10](#), [17](#), [22](#), [24](#)
- [9] H. Pehlivan, Y. Dalva, and A. Dundar, “StyleRes: Transforming the residuals for real image editing with StyleGAN,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1828–1837, 2023. [1](#), [7](#), [9](#), [10](#), [17](#), [34](#)
- [10] Y. Yu, L. Zhang, H. Fan, and T. Luo, “High-fidelity image inpainting with gan inversion,” in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pp. 242–258, Springer, 2022. [2](#), [10](#), [14](#), [17](#), [23](#), [24](#)
- [11] W. Wang, L. Niu, J. Zhang, X. Yang, and L. Zhang, “Dual-path image inpainting with auxiliary gan inversion,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11421–11430, 2022. [2](#), [10](#), [14](#), [17](#), [23](#), [24](#)
- [12] A. B. Yildirim, H. Pehlivan, B. B. Bilecen, and A. Dundar, “Diverse inpainting and editing with GAN inversion,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23120–23130, 2023. [2](#), [7](#), [10](#), [13](#), [14](#), [20](#), [23](#), [24](#), [25](#), [35](#), [36](#)
- [13] J. Li, N. Wang, L. Zhang, B. Du, and D. Tao, “Recurrent feature reasoning for image inpainting,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7760–7768, 2020. [4](#)
- [14] G. Liu, A. Dundar, K. J. Shih, T.-C. Wang, F. A. Reda, K. Sapra, Z. Yu, X. Yang, A. Tao, and B. Catanzaro, “Partial convolution for padding, inpainting, and image synthesis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. [4](#)
- [15] S. Zhao, J. Cui, Y. Sheng, Y. Dong, X. Liang, E. I. Chang, and Y. Xu, “Large scale image completion via co-modulated generative adversarial networks,” *arXiv preprint arXiv:2103.10428*, 2021. [4](#), [23](#), [24](#)

- [16] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, “MAT: Mask-aware transformer for large hole image inpainting,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10758–10768, 2022. 4, 5, 13, 23, 24, 25, 35
- [17] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, “Repaint: Inpainting using denoising diffusion probabilistic models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022. 4, 9, 23, 25, 27, 36
- [18] 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*, pp. 10684–10695, 2022. 4, 8, 9, 24, 25, 36
- [19] A. B. Yildirim, V. Baday, E. Erdem, A. Erdem, and A. Dundar, “Inst-inpaint: Instructing to remove objects with diffusion models,” *arXiv preprint arXiv:2304.03246*, 2023. 4
- [20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in Neural Information Processing Systems*, vol. 27, 2014. 5
- [21] 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*, pp. 4401–4410, 2019. 6
- [22] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, “Analyzing and improving the image quality of StyleGAN,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020. 6, 7, 9, 14, 25, 35
- [23] O. Patashnik, Z. Wu, E. Shechtman, D. Cohen-Or, and D. Lischinski, “Style-CLIP: Text-driven manipulation of StyleGAN imagery,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2085–2094, 2021. 6, 7, 9, 30, 36

- [24] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, “GANSpace: Discovering interpretable GAN controls,” *Advances in Neural Information Processing Systems*, vol. 33, 2020. 6, 7, 9
- [25] Y. Shen, J. Gu, X. Tang, and B. Zhou, “Interpreting the latent space of gans for semantic face editing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9243–9252, 2020. 7, 30, 36
- [26] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *Proceedings of the International Conference on Machine Learning*, pp. 2256–2265, pmlr, 2015. 8
- [27] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020. 8
- [28] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang, “Diffusion models: A comprehensive survey of methods and applications,” *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023. 8
- [29] G. Zhang, J. Ji, Y. Zhang, M. Yu, T. S. Jaakkola, and S. Chang, “Towards coherent image inpainting using denoising diffusion implicit models,” in *Proceedings of the International Conference on Machine Learning*, PMLR, 2023. 9, 23, 25, 27, 36
- [30] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, “Null-text inversion for editing real images using guided diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6038–6047, 2023. 9
- [31] O. Avrahami, O. Fried, and D. Lischinski, “Blended latent diffusion,” *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–11, 2023. 9
- [32] Y. Dalva and P. Yanardag, “Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24209–24218, 2024. 9, 31, 34

- [33] M. Brack, F. Friedrich, K. Kornmeier, L. Tsaban, P. Schramowski, K. Kersting, and A. Passos, “LEDITS++: Limitless image editing using text-to-image models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8861–8870, 2024. [9](#), [31](#), [34](#)
- [34] Y. Gao, F. Wei, J. Bao, S. Gu, D. Chen, F. Wen, and Z. Lian, “High-fidelity and arbitrary face editing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16115–16124, 2021. [9](#)
- [35] Y. Dalva, S. F. Altındış, and A. Dundar, “VecGAN: Image-to-image translation with interpretable latent directions,” in *European Conference on Computer Vision*, ECCV ’22, pp. 153–169, Springer, 2022. [9](#)
- [36] 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,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2287–2296, 2021. [14](#), [21](#), [22](#), [24](#)
- [37] Y. Alaluf, O. Patashnik, and D. Cohen-Or, “ReStyle: A residual-based StyleGAN encoder via iterative refinement,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6711–6720, 2021. [22](#), [24](#)
- [38] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *Advances in Neural Information Processing Systems*, vol. 30, 2017. [24](#)
- [39] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. [24](#)
- [40] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, “StarGAN v2: Diverse image synthesis for multiple domains,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8188–8197, 2020. [25](#)
- [41] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, “3D object representations for fine-grained categorization,” in *Proceedings of the IEEE International Conference on Computer Vision workshops*, pp. 554–561, 2013. [25](#)

- [42] A. Aksac, D. J. Demetrick, T. Ozyer, and R. Alhajj, “BreCaHAD: A dataset for breast cancer histopathological annotation and diagnosis,” *BMC Research Notes*, vol. 12, pp. 1–3, 2019. 25
- [43] S. Xu, Y. Huang, J. Pan, Z. Ma, and J. Chai, “Inversion-free image editing with natural language,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 31, 34

