

**ADVANCING BLIND FACE RESTORATION: ROBUSTNESS AND
IDENTITY PRESERVATION WITH INTEGRATED GAN AND CODEBOOK
PRIOR ARCHITECTURES**

SEYED NIMA TAYARANI BATHAIE

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

GRADUATE PROGRAM IN ELECTRICAL ENGINEERING AND COMPUTER
SCIENCE
YORK UNIVERSITY
TORONTO, ONTARIO

JANUARY 2024

© Nima Bathaie, 2024

Abstract

Blind Face Restoration (BFR) is a challenging task in computer vision, which aims to reconstruct High-Quality (HQ) facial images from Low-Quality (LQ) inputs. BFR presents as a challenging ill-posed problem, necessitating auxiliary information to constrain the solution space. While geometric and generative facial priors provide some support in BFR, their effectiveness wanes under intense degradation. Discrete codebook priors, though promising, grapple with the difficulty of associating intensely degraded images with their corresponding codes. To effectively address these limitations, this research introduces a two-stage restoration approach, termed Identity-embedded GAN and Codebook Priors (IGCP), which synergistically combines the strengths of both generative and codebook priors.

In the first stage, our approach employs a Generative Prior Restorer (GPR) network for initial image restoration. Distinct from existing methods that apply identity-based losses to the final restored image, our work innovates by embedding identity information directly into the style vectors of the StyleGAN2 network during the generation process. This is achieved through the introduction of an *identity-in-style* loss, ensuring superior fidelity and identity preservation even in severely degraded images

Proceeding to the second stage, the approach utilizes a two-component framework known as the Codebook Prior Restorer (CPR) network. This framework comprises a Vector Quantized AutoEncoder (VQAE) for artifact mitigation and to add a final touch of quality, complemented by introducing a Feature Transfer Module (FTM) that is demonstrated to be necessary to ensure fidelity and identity preservation.

Extensive experimental evaluations were conducted across five datasets, including our newly introduced CelebA-IntenseTest dataset. The results from these experiments demonstrate the remarkable efficacy of the IGCP approach. Notably, IGCP has shown exceptional performance in handling various degradation levels, setting new benchmarks in the domain of BFR.

Acknowledgements

I am deeply grateful to the individuals and organizations who have contributed to the completion of this thesis. Their support, guidance, and contributions have been invaluable.

First and foremost, I would like to express my heartfelt gratitude to my dedicated supervisor, Dr. Aijun An. Her unwavering support, insightful discussions, and collaborative brainstorming sessions have been instrumental in shaping this research.

I am also indebted to my esteemed committee members, Dr. James Harvey Elder and Dr. Gunho Sohn. Their constructive comments and invaluable feedback have significantly improved the quality of this thesis.

I extend my sincere appreciation to Dr. Hamidreza Dastmalchi for generously sharing his knowledge, which has greatly contributed to the advancement of this research.

Furthermore, I want to express my gratitude to Dr. Nastaran Babanezhad for her invaluable support that made all this possible.

I would like to extend my heartfelt thanks to my true friends, Dr. Javad Jafar Yahya and Dr. Zahra Dorostkar, for their optimism and encouragement throughout the journey of this dissertation.

My special thanks go to Sanat Chaudhry, who introduced me to the world of tennis, the best substitute activity when a project does not move forward!

I would also like to acknowledge the administrative and technical staff of our department and the Faculty of Graduate Studies, especially Ms. Nisa Lason, for their outstanding support throughout this journey.

In addition, I want to acknowledge the financial support provided by the NSERC CREATE Program: Smart Autonomous Robotic Technology for Space Exploration (SMART-ART), Grant No. CREATE 555425-2021, which partially funded this research.

Finally, I would like to extend my deepest gratitude to my family - my parents and my brother. The simple act of their existence and unwavering belief in me has been my greatest sources of encouragement and strength.

Table of Contents

Abstract	ii
Acknowledgements	iii
Table of Contents	iv
List of Tables	viii
List of Figures	ix
1 Introduction	1
1.1 Motivation	1
1.2 Applications	3
1.3 Problem Statement	8
1.4 Challenges	9
1.4.1 General Challenges	9
1.4.2 Face Domain Challenges	10
1.5 Objectives:	12
1.6 Contributions	13
1.7 Overview of the Thesis Organization	14
2 Literature Review	15
2.1 Introduction	15
2.2 Non-prior Based BFR	16
2.2.1 CNN Based BFR	16
2.2.2 GAN-Based BFR	18
2.2.2.1 GAN	18
2.2.2.2 Utilizing GANs for Face Restoration	21
2.2.3 Diffusion Based BFR	23
2.2.3.1 Denoising Diffusion Probabilistic Models (DDPM)	23
2.2.3.1.1 Forward Process	23

	2.2.3.1.2 Reverse Process (Backward Process)	24
	2.2.3.1.3 Objective Function	25
	2.2.3.2 DR2	26
2.3	Prior Based BFR	28
2.3.1	Geometric Prior Based BFR	28
2.3.2	Generative Prior Based BFR	30
2.3.2.1	Pretrained Generative Models	30
	2.3.2.1.1 StyleGAN	30
	2.3.2.1.2 BigGAN	31
	2.3.2.1.3 VQVAE	32
2.3.2.2	GAN Inversion	33
2.3.2.3	Optimization-Based GAN-Prior BFR	34
	2.3.2.3.1 PULSE	34
	2.3.2.3.2 Multi-code GAN prior	36
2.3.2.4	Learning-based GAN Prior BFR	38
	2.3.2.4.1 GPEN	39
	2.3.2.4.2 GLEAN	41
	2.3.2.4.3 GFP-GAN	43
2.3.3	Dictionary Prior Based BFR	45
2.3.4	Codebook Prior Based BFR	46
	2.3.4.1 VQFR	46
	2.3.4.2 CodeFormer	47
	2.3.4.3 RestorFormer	48
2.4	Datasets for BFR	49
2.4.1	CelebA	49
2.4.2	CelebAMask-HQ	50
2.4.3	HELEN	50
2.4.4	FFHQ	50
2.4.5	LFW	51
2.4.6	VGGFace	51
2.4.7	VGGFace2	52
2.5	Image Quality Assessment	54
2.5.1	PSNR	54
2.5.2	SSIM	55
2.5.3	MS-SSIM	56
2.5.4	LPIPS	57
2.5.5	FID	57
2.5.6	NIQE	58
2.5.7	MOS	59
3	Methodology	60
3.1	Motivation	60

3.1.1	Style-Identity Integration	61
3.1.2	Multi-resolution Feature maps Fusion	67
3.1.3	Codebook Prior Image Restorer	67
3.1.4	FTM Module	67
3.2	Model Architecture	67
3.2.1	GAN Prior Restorer (GPR, Stage I)	69
3.2.1.1	AutoEncoder (AE)	69
3.2.1.2	Identity embedded StyleGAN (IES)	70
3.2.2	Codebook Prior Restorer (CPR, Stage II)	71
3.2.2.1	Vector Quantized AutoEncoder (VQAE)	71
3.2.2.2	FTM Block	72
3.3	Training Scheme	74
3.3.1	GAN Prior Restorer (GPR, Stage I)	74
3.3.1.1	AutoEncoder (AE)	74
3.3.1.2	Identity embedded StyleGAN (IES)	75
3.3.2	Codebook Prior Restorer (CPR, Stage II)	77
4	Experiments	79
4.1	Datasets	79
4.1.1	Training Dataset.	79
4.1.2	Test Datasets.	80
4.1.2.1	Synthetic datasets	80
4.1.2.2	real-world datasets	83
4.2	Metrics	85
4.3	Implementation Details	85
4.4	Comparison with SOTA Methods	86
4.4.1	Quantitative Comparison.	86
4.4.1.1	Comparison on Synthesized Datasets	86
4.4.1.2	Comparison on Real-World Datasets	87
4.4.2	Qualitative Comparison	89
4.4.2.1	Visual Comparison on Synthesized Datasets	89
4.4.2.2	Visual Comparison on Real-World Datasets	92
4.4.3	Human Evaluation.	96
4.5	Ablation Study	99
4.5.1	GPR: Assessing Identity to Style Integration.	99
4.5.2	CPR: Significance of FTM Module.	101
4.6	Extension to other applications	102
4.6.1	Inpainting	102
4.6.2	Super-resolution	103
4.6.3	Colorization	105
5	Conclusion and future work	106

A	Appendix	108
A.1	StyleGAN	108
A.1.1	StyleGAN v1	109
A.1.1.1	Mapping Network	109
A.1.1.2	Synthesis Network	111
A.1.1.3	StyleGAN v1 Issues	112
A.1.2	StyleGAN v2	114
A.1.2.1	Weight Demodulation	114
A.1.2.2	Lazy Regularization	116
A.1.2.3	Path length regularization	117
A.1.2.4	Larger Network Architecture	117

List of Tables

2.1	Summary of Public Face Image Datasets for BFR.	53
3.1	Symbols used in the proof and their corresponding descriptions.	63
4.1	Quantitative comparison on CelebA-IntenseTest (heavy degradation).	86
4.2	Quantitative comparison on CelebA-Test (moderate degradation).	87
4.3	Quantitative comparison on real-world datasets.	87
4.4	Average ranks of each method in human evaluation.	96
4.5	Comparative Analysis of Intra-class and Inter-class Distances	101
A.1	Performance analysis of various mapping network depths	111

List of Figures

1.1	Face images illustrating various degradations from their original quality . . .	2
1.2	Application of BFR on movie restoration.	4
1.3	Real-world application of super-resolution in video conferencing	5
1.4	Digital photograph restoration.	6
1.5	BFR application on AI-generated Artworks	7
2.1	The basic architecture of a CNN.	16
2.2	GAN Framework.	19
2.3	Figure 1: Comparative Analysis of Loss Functions in GANs.	21
2.4	Sequential diffusion process in DDPMs.	23
2.5	Overview of the DR2 method. Image from [1]	27
2.6	The overall structure of PULSE.	35
2.7	Optimization-based GAN priors with multi-Code architecture	37
2.8	The architecture of GPEN method	40
2.9	Overview of GLEAN method	41
2.10	Overview of GFP-GAN framework	43
2.11	The VQFR framework.	47
2.12	CodeFormer framework	47
2.13	RestoreFormer framework	48
3.1	Maximizing mutual information for enhanced identity preservation	66
3.2	IGCP framework overview.	68
3.3	IGCP detailed architecture	73
3.4	IGCP training scheme	74
4.1	Example pairs of HQ ground-truth	80
4.2	CelebA-IntenseTest dataset	81
4.3	CelebA-Test dataset.	82
4.4	Representative examples from the LFW-Test dataset.	83
4.5	Representative examples from the WebPhoto-Test dataset.	84
4.6	Representative examples from the WIDER-Test dataset.	84
4.7	Qualitative Comparison of IGCP with SOTA models on CelebA-HQ dataset.	89

4.8	Qualitative comparison of IGCP with SOTA models on CelebA-IntenseTest.	90
4.9	Qualitative comparison of IGCP with SOTA models on CelebA-Test.	91
4.10	Qualitative comparison of IGCP with SOTA models on real-world datasets. .	92
4.11	Qualitative comparison of IGCP with SOTA models on WIDER-Test dataset.	93
4.12	Qualitative comparison of IGCP with SOTA models on WebPhoto-Test dataset.	94
4.13	Qualitative comparison of IGCP with SOTA models on LFW-Test dataset. .	95
4.14	Introduction page on the two surveys	97
4.15	Results from two user studies on image restoration.	98
4.16	Comparison of the style vectors generated by three variants of IGCP.	99
4.17	Comparisons of different GPR configurations.	100
4.18	Comparisons of different CPR configurations.	101
4.19	IGCP face inpainting results.	102
4.20	IGCP super resolution results (I).	103
4.21	IGCP super resolution results (II).	104
4.22	IGCP colorization results.	105
A.1	Mapping network in StyleGAN v1	110
A.2	StyleGAN v1 synthesis network	112
A.3	blob-like artifacts in In StyleGAN v1	113
A.4	Phase artifacts resulting from progressive growing in StyleGAN v1	114
A.5	Revised architecture in StyleGAN2.	115
A.6	Replacement of instance normalization with a “demodulation” operation . . .	116
A.7	StyleGAN v2 network size	118

Chapter 1

Introduction

1.1 Motivation

In our digitally-driven world, the significance of facial images extends far beyond simple visual representation; they play pivotal roles in areas ranging from social media interactions to biometric identification. However, face images captured under unconstrained conditions, commonly termed *in the wild*, are often marred by a plethora of degradations. Such imperfections can range from compression artifacts and noise to blurriness and low resolution. This is especially true when the subjects of these images are located far away from the capturing cameras.

To counteract these challenges, the domain of image processing has explored various strategies, with machine learning techniques distinguished as a cost-efficient tool for image restoration. Face restoration, a sub-domain of image restoration, is the process of transforming a Low-Quality (LQ) face image into a high-quality (HQ) version.

Depending on the nature of the degradation, **face restoration** can be branched into five principal categories (see Figure 1.1):

1. **Face Denoising**, which involves removing noise, such as Gaussian noise, from the face image[2, 3, 4].
2. **Face Deblurring**, which focuses on retrieving a clear face from a blurry one, often due to factors like camera shake or moving subjects [5, 6, 7].

3. **Face Super-Resolution (FSR)**, Often called face hallucination [8], its goal is to improve the quality and clarity of low-resolution (LR) facial images and produce high-resolution (HR) counterparts. [9, 10, 11, 12, 13, 14].
4. **Face Artifact Removal**, which addresses the recovery of HQ images from those affected by compression artifacts during storage or transmission[15, 16].
5. **Blind Face Restoration (BFR)**, which aims to restore HQ images from their degraded LQ versions without specific information on the type of degradation[16, 17, 18, 19, 1].



Figure 1.1: Face images illustrating various degradations from their original quality: (a) HQ original, (b) noise degradation, (c) blur, (d) visual artifacts, (e) low resolution, and (f) a mix of these degradative factors.

BFR stands out for its capability to restore HQ face images from degraded versions, all without prior knowledge of the nature of degradation. It grapples with complexities arising from random combinations of noise, blurriness, low resolution, and compression artifacts. The objective of BFR is not merely to enhance image quality but to regenerate an image that aligns closely with the original HQ ground-truth, ensuring the preservation of intricate details and inherent facial features. In practice, it is typically observed that both the types of degradation and the degradation parameters remain unknown beforehand. Therefore, BFR merits increased attention.

Despite the growing interest in BFR within the image processing and computer vision communities, a significant challenge persists: *preserving identity information during restoration*. This thesis centers on the ambition to develop a method that addresses this challenge. By evaluating and building upon current methodologies, this research aims to elevate the effectiveness of BFR, ensuring the accurate representation of identity in restored images.

1.2 Applications

The ability to produce HQ facial images is crucial for numerous practical applications that contribute to the growing popularity of BFR. The following are some of the key applications of BFR:

- **Surveillance:** In the context of security and public safety, HQ facial images are crucial for accurately identifying individuals and monitoring their activities. BFR can enhance the quality of images captured by surveillance cameras, even in challenging conditions such as low light or at a distance, making it easier for law enforcement and security personnel to recognize and track individuals.
- **Biometrics:** Biometric systems, such as facial recognition, are widely used in various domains and scenarios, where accuracy and reliability are crucial factors. However, the quality of the images used for identification can be affected by several factors, such as LQ cameras or large distances between the subject and the camera, which can degrade the performance of these recognition systems. Therefore, some researchers have focused on incorporating BFR techniques into facial recognition methods, aiming to overcome these difficulties and improve recognition accuracy [20, 21, 22, 23].
- **Social Media and Photography:** In the near future, most smartphones are likely to incorporate neural network-based photo upscaling, with BFR playing a significant role. BFR techniques can be applied to enhance the visual quality of images captured by smartphones and shared on social media platforms. This can lead to more appealing and visually striking content, making it more likely to capture attention and garner engagement from users. Digital zooming in cameras is another example where BFR can be useful. When zooming in on an image captured by a camera, e.g. a smartphone, the image quality can degrade, resulting in blurry or pixelated images. By using BFR techniques, the visual quality of zoomed-in facial images can be enhanced, resulting in clearer and more detailed images.
- **Entertainment:** In movies, video games, and virtual reality, realistic and HQ facial images are essential for creating immersive and engaging experiences. BFR techniques can be employed to enhance the visual quality of facial images in these applications, resulting in more lifelike and compelling characters.



Figure 1.2: Application of BFR on movie restoration: (a) faces in the movie restored using the method in [12], (b) face in the movie restored with the method proposed in [18] method.

- Video Conferencing:** In recent years, video conferencing and remote work have become increasingly important in many industries. These applications often grapple with challenges like noise and LR webcam images that can make it difficult to identify facial features and expressions. Image restoration has emerged as a solution to this problem, enabling the enhancement of LQ images and making virtual meetings more engaging and interactive. Consequently, a substantial body of research has been devoted to improving the performance of web-streaming through the application of super-resolution upscaling techniques [24, 25, 26]. Simultaneously, several companies, such as Topaz Labs¹, AVC Labs², and Nvidia³, offer AI-based upscaling solutions for web-streaming. For example, Nvidia’s popular cloud-based platform, Nvidia Maxine, utilizes FSR and other AI-based upscaling techniques to improve the quality of video conferencing and remote work applications.

With features such as virtual backgrounds and noise cancellation, Nvidia Maxine enhances the virtual meeting experience while reducing bandwidth usage. Additionally, advanced AI algorithms can detect and enhance facial features, gestures, and expressions, further improving the quality of virtual interactions (figure 1.3).

¹ <https://www.topazlabs.com/topaz-video-ai>

² <https://www.avclabs.com>

³ <https://developer.nvidia.com/maxine>

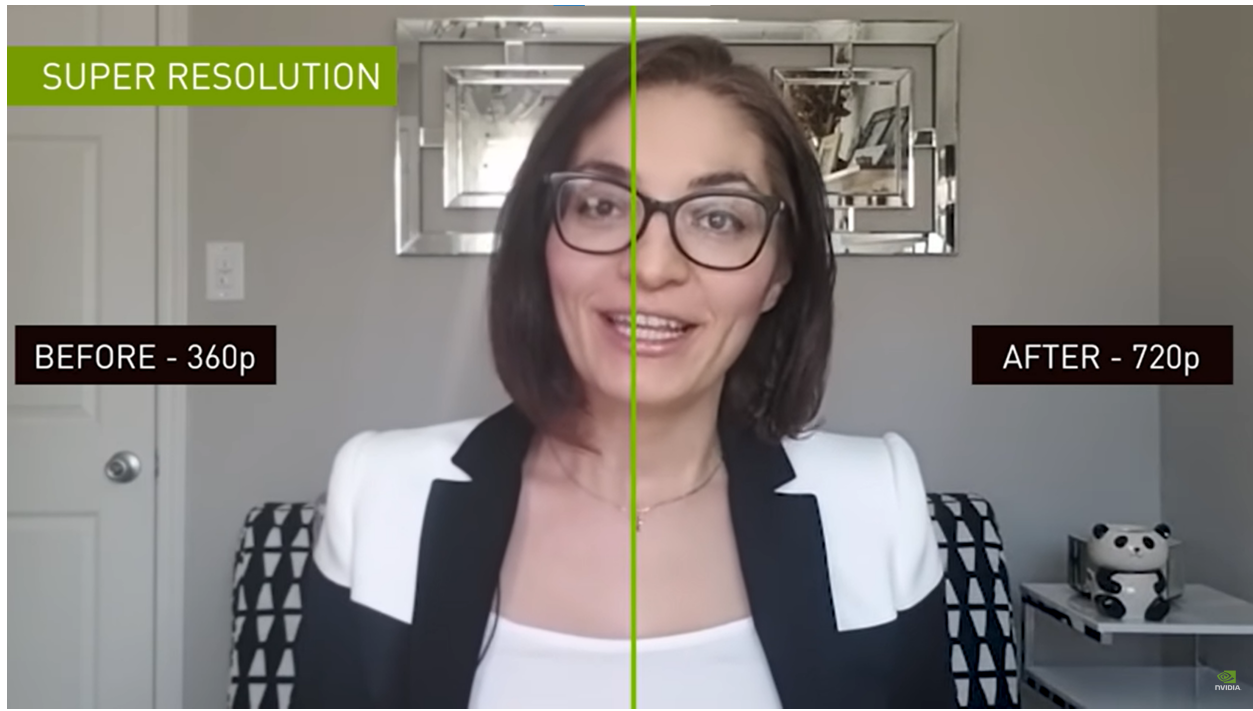


Figure 1.3: Real-world application of SR in video conferencing: Nvidia’s Maxine, an AI-driven platform enhancing the quality and user experience of real-time video-conference applications, where SR allows participants to transmit LR video streams, which are then upscaled on the server, reducing bandwidth demands and promoting stable performance in areas with inconsistent network connectivity.

- **Old Photo Restoration:** BFR techniques can be applied to old or damaged photographs, improving their quality and revealing details that may have been lost due to aging, damage, or degradation. This application can be useful for historians, genealogists, and individuals looking to preserve family memories.



Figure 1.4: Digital photograph restoration: Old historic photo of the 5th Solvay Conference from 1927, restored using the BFR method proposed in [18]. Faces individually cropped, aligned, enhanced, and pasted back onto the original photo for a remarkable transformation.

- **Refinement of AI-generated Artworks** BFR can also be employed to enhance faces in AI-generated images. Figure 1.5 illustrates two instances generated using Stable Diffusion[27], along with the improved outcomes achieved through CodeFormer[18].



Figure 1.5: BFR application on AI-generated Artwork: (a) Image generated with Stable Diffusion model [27]. (b) Enhanced output achieved through CodeFormer[18]

1.3 Problem Statement

Real-world facial images can encounter multiple types of degradation during their capture, transmission, and storage phases. The degradation may take several forms, such as noise, blur, and compression artifact. We can represent this degradation process using a function Φ , which maps a HQ face image $\mathbf{h}$ to its degraded counterpart $\mathbf{l}$, given by:

$$\mathbf{l} = \Phi(\mathbf{h}; n_\delta) \quad (1.1)$$

where n_δ denotes additive white Gaussian noise with a noise level δ . In practice, only the degraded image $\mathbf{l}$ is available, and no information about the degradation process or its parameters is known. The degradation function in real applications depends on many uncontrolled conditions, including environmental factors, lens specification, motion, and blur. Face restoration aims to reverse the effects of degradation and produce an approximation of the original face image $\hat{\mathbf{h}}$:

$$\hat{\mathbf{h}} = \Phi^{-1}(\mathbf{l}; n_\delta) \quad (1.2)$$

where Φ^{-1} represents the face restoration model. If the degradation factors are provided, the face restoration task is termed non-blind face restoration, which includes face denoising, face deblurring, FSR, and face artifact removal. Otherwise, the face restoration task is regarded as BFR, in which the degradation process is defined as a combination of blur, noise, JPEG compression artifacts, and low resolution as follows [28, 16, 17]:

$$\mathbf{l} = \left\{ [\mathbf{h} \otimes k \downarrow_r + n_\delta]_{\text{JPEG}_q} \right\} \uparrow_r \quad (1.3)$$

where k denotes the blurring kernel, $\otimes$ represents convolution, $\downarrow_r$ symbolizes the down-sampling operation with factor r , n_δ is the additive Gaussian noise, and $[\]_{\text{JPEG}_q}$ implies JPEG compression with a quality factor q . $\uparrow_r$ signifies the up-sampling operation with a scaling factor r .

In BFR, the aim is to recover HQ face images by modeling the inverse process of the above degradation model.

1.4 Challenges

In order to provide a better demonstration, we categorize the challenges associated with BFR into two groups: *General Challenges* encompassing issues that are common to Blind Image Restoration (BIR) in general, and *Face Domain-Specific Challenges* that focuses on obstacles specific to the face domain.

1.4.1 General Challenges

BFR is a sub-domain of BIR, focusing specifically on enhancing the quality of facial images. Consequently, the challenges inherent in BIR are also applicable to BFR tasks. In this section, we outline the challenges associated with the broader domain of BIR, which have a direct impact on the performance and effectiveness of BFR techniques. The following are some of the key challenges in BIR:

- **Ill-posed problem:** BIR presents an ill-posed inverse problem. In mathematical terms, a problem is termed 'ill-posed' when it does not meet the three Hadamard criteria of existence⁴, uniqueness⁵, and stability⁶ [29]. Specifically, in the context of BIR, this means that from a given LQ input, multiple HQ images could be reconstructed, making it indeterminable as to which one is the true HQ representation. Consequently, it becomes impossible to pinpoint the correct ground-truth output in practice. One potential solution to mitigate this issue is to generate an average of all possible HQ images that downscale to the same LQ input. However, this approach can lead to a loss of sharpness, especially in textures and edges—referred to as 'high-frequency details'—resulting in a blur in the final image.
- **Degradation factors:** Real-world LQ images are often affected by various degradation factors such as noise, blur, compression artifacts, and camera limitations. These degradations can further complicate the BIR process and affect the quality of the reconstructed HQ images. Developing BIR techniques that are robust to such degradations and can effectively handle these challenges is crucial for achieving optimal performance.
- **Computational Complexity:** Advanced BIR techniques often involve complex operations such as feature extraction, mapping, and reconstruction. These operations can be computationally expensive, particularly when processing large-scale datasets or real-time applications. Designing BIR algorithms that can deliver HQ results while maintaining computational efficiency is an essential challenge to address.

⁴ Existence: A solution to the problem exists.

⁵ Uniqueness: There is only one unique solution to the problem.

⁶ Stability: Small changes in the input lead to small changes in the output (the solution is not overly sensitive to changes in the input).

- **Loss of fine details and textures:** Restoring fine details and textures in a HQ image is difficult due to the limited information available in the LQ input. Many BIR methods tend to generate overly smooth images, resulting in a loss of details and textures.
- **Training data limitations:** Training data for BIR algorithms typically consists of HQ images from which their LQ counterparts are obtained using a degrading function, such as in Equation 1.3. However, acquiring such HQ training data can be challenging due to issues such as copyright restrictions, the time-consuming process of data collection, and ensuring diverse image scenarios and lighting conditions. Furthermore, the performance and robustness of BIR methods depend not only on the quantity of data but also on its quality and variety. Therefore, constraints in data availability can significantly influence the performance and generalizability of the developed BIR algorithms.
- **Evaluation metrics:** Selecting appropriate evaluation metrics for BIR is challenging, as traditional metrics like PSNR and SSIM [30] may not always correlate with human perception of image quality. Recently, researchers have proposed perceptual-based metrics, but there is still no consensus on the best metrics to use.
- **Trade-off between fidelity and perceptual quality:** There is often a trade-off between the objective fidelity of the reconstructed image (e.g., measured by PSNR) and its subjective perceptual quality. Striking the right balance is challenging, and an algorithm that performs well in one aspect may not necessarily perform well in the other.
- **Robustness to real-world images:** Many BIR algorithms are trained and evaluated on synthetic data, where the LQ images are derived from their HQ counterparts using a defined degradation function. However, this approach may not accurately capture the complexity of real-world images. Ensuring that algorithms generalize well to real-world scenarios remains a significant challenge.

1.4.2 Face Domain Challenges

Despite significant advancements in BFR techniques, several challenges persist, making it difficult to achieve optimal results. Overcoming these challenges will contribute to the advancement of the BFR field and unlock new possibilities for a wide range of applications. In this section, we discuss the primary challenges in BFR.

- **Complex and Sensitive Facial Structure:** Faces consist of various features such as eyes, nose, mouth, and contours, each with their unique details and textures. The accurate reconstruction of these features is crucial for maintaining the identity and natural appearance of the subject. Preserving fine details while avoiding artifacts and distortions is a challenging task in BFR.

- **Variability in Pose, Expression, and Illumination** Facial images can exhibit a wide range of variability in pose, expression, and illumination. These factors add complexity to the BFR process, as they affect the appearance of facial features and the overall structure of the face. Developing BFR techniques that can effectively handle such variability and produce visually consistent and accurate results is a significant challenge.
- **Limited Ground-Truth Data** The availability of HQ ground-truth data is crucial for training and evaluating BFR algorithms. However, obtaining perfectly aligned and matched pairs of HQ and LQ facial images is difficult, particularly when considering the variability in pose, expression, and illumination. The limited availability of ground-truth data makes it challenging to train robust BFR models and accurately evaluate their performance.
- **Loss of Identity Information** A critical aspect of BFR is preserving the identity information of the subject while enhancing the quality the facial image. Some BFR techniques may generate visually appealing results but fail to maintain the subject's identity, leading to reduced effectiveness in applications such as face recognition and biometrics. Developing BFR algorithms that can effectively balance visual quality and identity preservation remains a challenging task.
- **Robustness to Noise and Degradations** Distorted facial images are often affected by various degradations, such as noise, blur, and compression artifacts. These degradations can further complicate the BFR process and affect the quality of the reconstructed HQ facial images. Developing BFR techniques that are robust to such degradations and can effectively handle these challenges is crucial for achieving optimal performance.

The challenges outlined above not only highlight technical limitations in BFR but also underscore significant risks in certain applications, particularly where system reliability is paramount. In such contexts, BFR can inadvertently cause identity mismatches between LQ and HQ images, potentially leading to false positives. This issue often stems from generative models' propensity to interpolate from training data, creating features not present in the original image. Ensuring reliability in critical applications like legal or security contexts thus becomes a complex challenge.

Moreover, these technical hurdles intertwine with ethical considerations. The dependence on AI for crucial decisions involving facial recognition necessitates a comprehensive understanding of the model's limitations and the potential consequences of its inaccuracies. It is imperative to approach the integration of BFR in sensitive applications with caution, prioritizing the development of robust models and ethical frameworks that consider the full spectrum of issues that might arise from errors in recognition grounded in Artificial Intelligence.

1.5 Objectives:

In this research endeavor, our primary aim is to develop an innovative BFR method that is robust to severe degradation, excels in preserving the identity, and surpasses the capabilities of existing State-of-the-Art (SOTA) techniques. To achieve this overarching goal, we have outlined the following specific objectives:

1. Design and implement a BFR method that is robust to severe degradation, and capable of preserving identity information during the restoration process, ensuring that the individual's unique features remain recognizable in the reconstructed HQ images.
2. Evaluate the proposed method on various datasets, namely characterized by different degrees and types of degradation. This will provide a broad perspective on the method's effectiveness across different challenges.
3. Investigate the qualitative performance of the proposed method in reconstructing HQ faces with photo-realistic details from severely degraded images in the wild, emphasizing its utility in real-world applications.
4. Assess the effectiveness of the proposed method in terms of key parameters like identity preservation, fidelity, and overall image quality. This will involve comparing the performance of the proposed method with other SOTA methods, using a set of recognized quantitative metrics.
5. Conduct human evaluations with the aim to capture subjective feedback on the images produced by both SOTA methods and our proposed method.
6. Dissect and evaluate the contribution of individual components or configurations within the broader framework of the proposed method.
7. Explore the method's versatility by adapting it to various face manipulation tasks beyond restoration.

1.6 Contributions

In pursuit of our research objectives, we have made the following contributions:

1. To achieve the first objective, We proposed a novel two-stage BFR method designed to leverage the complementary strengths of two types of prior networks. The continuous latent space of GAN (Generative Adversarial Network) prior networks was utilized to address uncertainties related to identity, while the discrete space of codebook prior networks was employed to manage ambiguities associated with severe degradations.
2. In pursuit of the second objective, we conducted comprehensive evaluations on diverse datasets, both synthetic and real-world, with varying levels of degradation, demonstrating the method’s adaptability and robustness in addressing different challenges. The datasets included are real-world datasets LFW-Test [16] with low-level degradation, WebPhoto-Test [16] with mid-level degradation and WIDER-Test [18] with high level degradation. And two synthetic datasets, CelebA-Test proposed by [16] with mid-level degradation, And our propose CelebA-IntenseTest dataset with a high level degradation.
3. To achieve the third objective, we showcased the method’s practical applicability in restoring highly degraded facial images, resulting in photo-realistic outcomes suitable for real-world scenarios.
4. To assess the effectiveness (Objective 4), we compared our method with state-of-the-art approaches using quantitative metrics such as PSNR, SSIM [30], FID [31], LPIPS [32]. Additionally, we evaluated identity preservation by employing the cosine similarity of features extracted from the ArcFace network [33], denoted as IDS.
5. To capture diverse user perspectives and feedback (Objective 5), we conducted two user studies, collecting valuable subjective insights through an online survey, which contributed to the validation of our approach’s practicality.
6. Achieving Objective 6, we conducted an ablation study that systematically analyzed the contributions of individual components within our method, providing insights into its inner workings and areas for further refinement.
7. In line with Objective 7, we extended the versatility of our method to various face manipulation tasks beyond restoration, including super-resolution, image coloring, and inpainting, broadening its scope and potential applications in the field of image processing.

1.7 Overview of the Thesis Organization

The remainder of the thesis is structured as follows:

1. In the **Literature Review** chapter, we delve into the rich history and development of BFR methods. We discuss a wide range of techniques, from CNN-based methodologies to the more recent advancements using GANs and GAN priors. By emphasizing the strengths and shortcomings of each method, we provide context and motivation for our proposed BFR approach.
2. The **Methodology** chapter is devoted to a detailed exposition of the proposed method. It begins with the motivations and inspirations behind the key components of our approach. The chapter delves into the distinct modules that collectively form the backbone of our solution. Additionally, we explore the architecture of our model, detailing how different components interact and work in tandem. Lastly, the training schemes, including the individual training processes for each stage of the model, are elaborated upon.
3. The **Experiments** chapter brings the methodology to life, presenting a comprehensive evaluation of our approach. Starting with a description of the datasets utilized – both for training and testing – we subsequently delve into the metrics used to assess performance. Comparative studies with existing SOTA methods are shown, providing both quantitative and qualitative insights. Human evaluations and ablation studies reveal more nuanced aspects of our model’s performance. Furthermore, we also demonstrate the versatility of our approach by extending its applications beyond restoration to tasks like inpainting, super-resolution, and colorization.
4. In the **Conclusion and Future Work** section, we summarize the contributions and findings of our research. We also explore the avenues for future investigations and enhancements. This segment aims to inspire subsequent research endeavors by suggesting potential improvements, extensions, and broader applications for our proposed methodology.

Chapter 2

Literature Review

2.1 Introduction

General image restoration methodologies focus on devising effective techniques for the recovery of sharp, natural images. However, given the unique and highly structured nature of the human face, there are specific characteristics often overlooked by these generic methods. This realization has driven the development of face restoration methods that incorporate face-specific prior knowledge, striving for more accurate reconstructions of facial structures. Such face-specific priors primarily arise from the understanding that human facial features exhibit minimal variations in controlled environments.

Face restoration has a well-established history in the domain of computer vision. In 2000, Baker and Kanade [34] introduced the concept of FSR. Their pioneering work used a multi-level learning and prediction model based on the Gaussian image pyramid, aimed at enhancing the resolution of LR facial images. The success and potential of this approach prompted further exploration, leading to a series of machine learning-based approaches targeting FSR challenges [35, 36, 37, 38, 39, 40].

As the field progressed, "Blind Face Restoration" (BFR) emerged as a significant area of research and gained prominence towards the end of the 2010s. This surge in attention can be attributed to breakthroughs in deep learning and the increasing power of computational methodologies in image processing. Unlike its predecessors, BFR seeks to restore facial images without specific knowledge of the degradation type. This approach offers a broader application range than FSR and encompasses additional challenges such as face deblurring, denoising, and compression artifact removal.

Building on these foundational concepts, face restoration methods are broadly categorized into two groups: prior-based deep restoration methods and non-prior based deep learning approaches. Within prior-based methodologies, a further sub-categorization emerges, consisting of geometric prior, reference prior, generative prior and codebook prior-based deep restoration techniques. Each of these methods will be discussed in detail in the subsequent sections.

2.2 Non-prior Based BFR

2.2.1 CNN Based BFR

Convolutional Neural Networks (CNNs) are popular tools in deep learning, particularly for tasks involving images. Like Deep Neural Networks (DNNs), CNNs consist of an input layer, multiple hidden layers, and an output layer. The distinctive feature of CNNs lies in their hidden layers, which incorporate specific layers known as convolutional and pooling layers. These layers allow the network to detect and interpret spatial features and patterns in images, which is vital for tasks such as image classification and image restoration. The basic architecture of a CNN is illustrated in Figure 2.1.

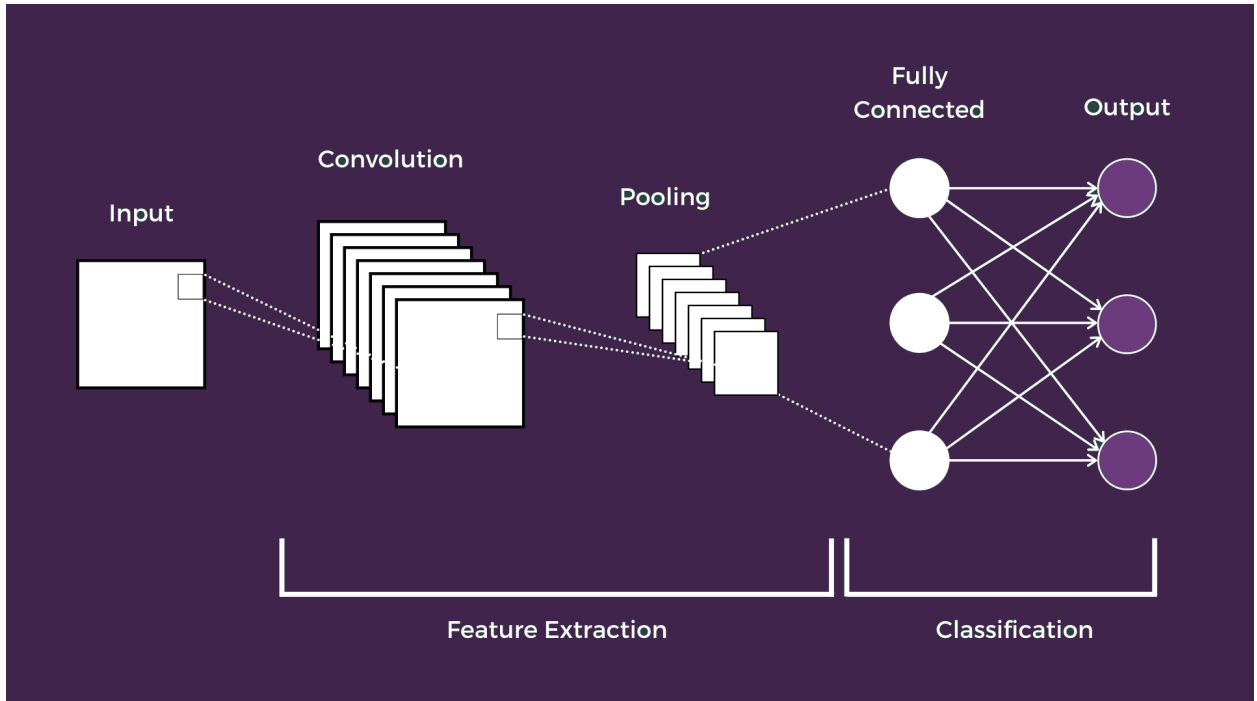


Figure 2.1: The basic architecture of a CNN. A CNN is composed of multiple layers such as convolutional, pooling, and fully connected ones. Within the convolutional layer, filters are applied to the input image, capturing features like edges, corners, and textures. Subsequently, the pooling layer reduces the size of the feature maps, retaining essential details. The fully connected layer then employs these features to deduce attributes of the input image.

The prominence of CNNs in the realm of deep learning can be credited to Yann LeCun. His innovative application of a CNN architecture, known as LeNet-5 [41], demonstrated remarkable accuracy in classifying hand-written digits. LeNet-5 was designed to automatically identify and adapt to pixel-level features within the training images. LeCun’s achievement with LeNet-5 captured the attention of the wider Computer Vision community. This sparked an intensive exploration to design enhanced CNN architectures.

By 2012, with the rise of deep learning’s influence and popularity in computer vision [42], CNNs became the primary choice for feature extraction and learning patterns in image data. Inspired by the pioneering deep learning general image super-resolution method, SRCNN [10], researchers have proposed incorporating CNN networks into the FSR and BFR tasks. In the subsequent years, a diverse range of CNN models and architectures emerged to improve the learning capability of CNNs. The VDSR [11] method uses a deeper CNN network to improve the results and introduces the concept of residual learning to increase training convergence and prevent gradient vanishing. Authors in [43, 44] adopted Recursive Neural Networks (RNNs) to share parameters between different layers and achieved similar results.

While generic super-resolution CNNs can be used to enhance facial images, target-specific networks that are specifically trained on facial images tend to produce more accurate results. Researchers in [45] developed a Bi-Channel Convolutional Neural Network (BCCNN) to directly learn a mapping from LQ images to the corresponding HQ ones. [46] introduced the Iterative Back Projection method for FSR network and achieved convincing results. Authors in [47] proposed a Residual Back-Projection Network (RBPNET) to learn the residual information between the reconstructed image and the ground truth one by gradually updating the network over multiple steps. This allows the network to gradually improve the quality of the reconstructed image, resulting in a more accurate representation of the HR face image. In [48], the authors presented a novel global-local fused network (GLFSR) to reconstruct the high-frequency details of facial images. Similarly, the Dual-Path Deep Fusion Network (DPDFN) [49] employs two separate branches to learn both global facial contours and local facial component details and combines the results from these two branches to generate the final HR image.

Although CNN-based face restoration techniques provide superior results compared to traditional methods, they yield blurry outputs which lack fine details. This is primarily due to the use of pixel-wise loss function for training these networks. To mitigate this problem, researchers proposed to combine an adversarial loss term in GAN frameworks.

2.2.2 GAN-Based BFR

2.2.2.1 GAN

The Generative Adversarial Network (GAN) was introduced by Goodfellow et al. [50, 51]. It represents a deep generative model designed for unsupervised learning. The GAN architecture includes two distinct networks: the generator, denoted as G , and the discriminator, denoted as D . The generator's function is to take a random noise, z , and output an approximation of real data, represented as $G(z)$. Conversely, the discriminator's primary function is to map input data x to a probability value p such that $D : x \rightarrow p$ where $p \in [0, 1]$. This probability indicates whether the input data originates from the actual dataset (with a probability close to 1) or from the synthetic data $G(z)$ (with a probability approaching 0). The training process alternates between optimizing G and D . After adequate training, if G produces data that aligns closely with the real data distribution, D will find it challenging to differentiate effectively being "fooled" by G . The GAN pipeline is visualized in Figure 2.2.

In standard GAN formulations, while D attempts to differentiate real from synthesized data, G aims to deceive D . This adversarial relationship can be mathematically described by the following minimax game:

$$\min_G \max_D \mathbb{E}_{x \sim P_r} [\log D(x)] + \mathbb{E}_{x \sim P_g} [\log (1 - D(x))] . \quad (2.1)$$

The associated loss functions for G and D are defined as:

$$\mathcal{L}_G = -\mathbb{E}_{x \sim P_g} [\log(1 - D(x))] , \quad (2.2)$$

and

$$\mathcal{L}_D = -\mathbb{E}_{x \sim P_r} [\log D(x)] - \mathbb{E}_{x \sim P_g} [\log(1 - D(x))] , \quad (2.3)$$

respectively, where P_r and P_g denote the distributions of real and generated data.

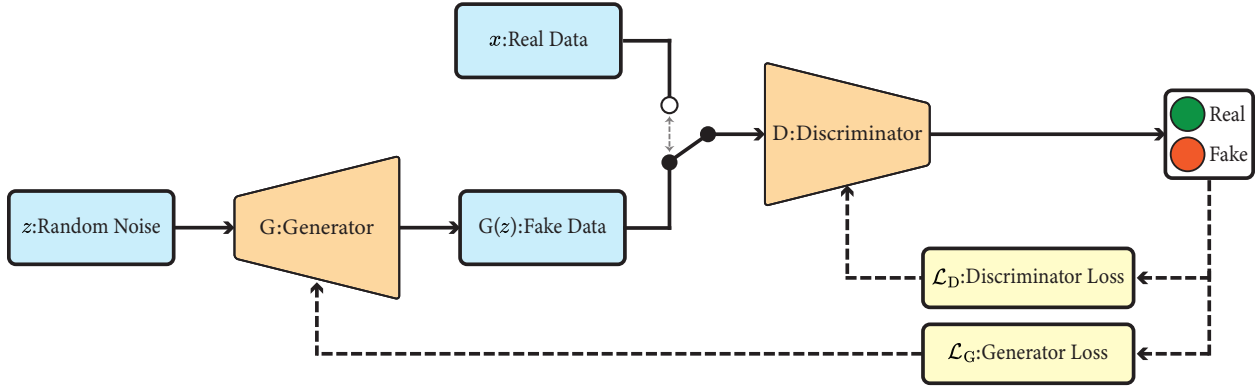


Figure 2.2: GAN Framework.

In the early stages of GAN training, the real data distribution and the generated data distribution are usually very different. This means that even if D is not well trained, it can still distinguish between real and generated data easily. Therefore, $D(x)$ is close to 1 for real data and near 0 for generated data. According to equation 2.2, this leads to the gradient $\frac{\partial \mathcal{L}_G}{\partial D(x)}$ becoming very large, which causes substantial changes in the generator's parameters.

For a specific sample x , irrespective of whether it originates from the real data distribution or the generated distribution, the loss contribution from this sample, based on equation 2.3, is:

$$\mathcal{L}_D = -P_r(x) \log D(x) - P_g(x) \log(1 - D(G(x))). \quad (2.4)$$

From this, the optimal discriminator can be derived as:

$$D^*(x) = \frac{P_r(x)}{P_r(x) + P_g(x)}. \quad (2.5)$$

Drawing from equation 2.5, after sufficient training, if the generated data distribution $P_g(x)$ aligns closely with the real data distribution $P_r(x)$ such that $P_g(x) = P_r(x)$, the output probability $D(x)$ from the discriminator will converge to 0.5. Referring back to equation 2.2, at this point, the gradient for the generator will approach zero: $\frac{\partial \mathcal{L}_G}{\partial D(G(z))} \rightarrow 0$. Consequently, updates to the generator cease, ensuring no further reduction in the generator's loss.

In brief, classical GANs exhibit certain limitations. A well-trained discriminator results in a vanishing gradient for the generator, halting its improvement. On the other hand, an under-trained discriminator leads to imprecise gradients, causing the generator to oscillate inconsistently.

Addressing these issues, enhanced versions of GANs have been proposed [52, 53, 54]. The Wasserstein GAN (WGAN) [52] employs the Earth-Mover (EM) distance as a metric between the real and generated data distributions. In contrast to the Jensen–Shannon (JS) divergence and KL divergence utilized in traditional GANs, the Wasserstein distance offers a solution to a fundamental problem. Specifically, in instances where the generated and real distributions either don’t overlap or the overlapping regions are negligible, traditional GANs can be highly unstable. The Wasserstein distance, however, can accurately reflect the actual distance even when the distributions lack overlap.

Typically, the generator network G in WGANs seeks to minimize the loss function in equation 2.2; simultaneously, the discriminator D of WGANs aims to minimize the loss function in equation 2.3.

The Least Squares Generative Adversarial Network (LSGAN) [53] offers an innovative solution to a problem identified in classical GANs. The authors of LSGAN contended that utilizing sigmoid cross-entropy loss in classical GANs can potentially lead the generator to produce LQ images, even if it has successfully confused the discriminator.

To address this issue, LSGAN adopts least squares loss as the objective function, aiming to guide the generated (fake) samples to resemble real data samples more closely. This approach significantly differs from the sigmoid cross-entropy loss employed in traditional GANs, and this difference in behavior becomes evident during the training of the GANs.

LSGAN’s approach of replacing sigmoid cross-entropy loss with least squares loss aims to direct fake samples closer to the decision boundary, thus guiding the generator more effectively. LSGAN’s least squares loss function penalizes samples that are on the correct side of the decision boundary but are still far from the real data. This penalization is more effective than the sigmoid cross-entropy loss function in guiding the generator to produce samples that are not just on the correct side of the boundary but are also closer to the real data. This method generates more gradients for updating the generator, leading to more stable training and higher-quality images. In essence, LSGAN enhances the ability of the generator to produce samples that align more closely with the decision boundary, thereby potentially improving the efficiency of model optimization.

The generator (G) of LSGAN tries to minimise

$$L_g = \frac{1}{2} \mathbb{E}_{x \sim p_g} [(D(G(x)) - c)^2], \quad (2.6)$$

while the discriminator D strives to minimise the loss function

$$L_d = \frac{1}{2} \mathbb{E}_{x \sim p_r} [(D(x) - b)^2] + \frac{1}{2} \mathbb{E}_{x \sim p_g} [(D(G(x)) - a)^2]. \quad (2.7)$$

The values of a , b , and c are subject to the conditions $b - c = 1$ and $b - a = 2$. For instance, one could choose $a = -1$, $b = 1$, and $c = 0$.

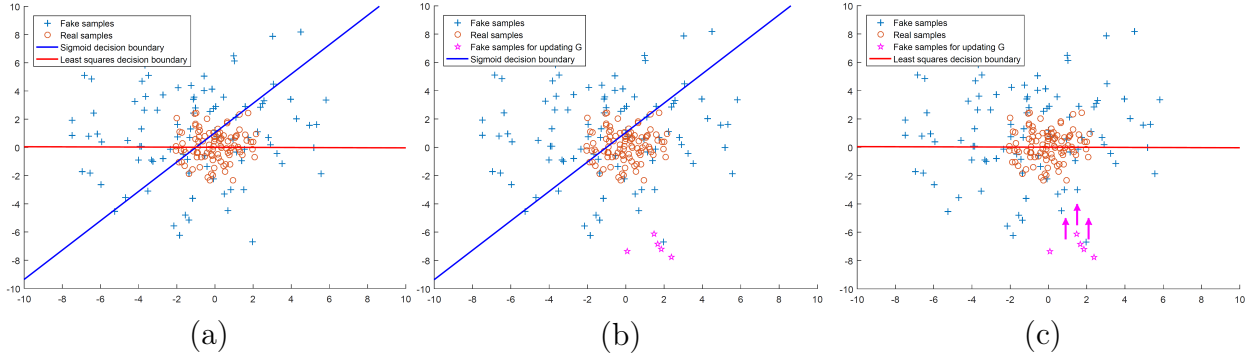


Figure 2.3: Figure 1: Comparative Analysis of Loss Functions in GANs. (a) Visualizes the decision boundaries for sigmoid cross-entropy and least squares loss functions. (b) Demonstrates the sigmoid cross-entropy loss function’s decision boundary, where it produces very small errors for fake samples (in magenta) that are correctly classified yet distant from the boundary. (c) Illustrates the least squares loss function’s decision boundary, which actively penalizes fake samples (in magenta), thereby compelling the generator to produce samples that gravitate towards the decision boundary, enhancing overall GAN performance. Image from [53]

2.2.2.2 Utilizing GANs for Face Restoration

GANs can generate trustful images from noise inputs. Therefore, one idea is to use GANs for facial image reconstruction. In the evolving landscape of GAN-based Face Restoration (FR) methods, various innovative approaches have been proposed. In the initial stages, Yu et al. [55] introduced URDGN, a method which utilized a generative model to synthesize restored

face images, and a discriminative model to differentiate between real HR face images and artificially super-resolved outputs. This was a seminal work in using GANs for ultra-resolving face images.

MLGE [56] is another successful FRS technique based on GAN networks. The novelty of MLGE is that the generator synthesizes edge maps along with high-level feature maps to reconstruct the HR image. The discriminator network, on the other hand, discriminated not only the image but also the edge maps. This marked a step forward in the utilization of GANs for face restoration, focusing on enhancing image quality through advanced reconstruction methods.

Xu et al. [57] introduced a method based on Generative Adversarial Networks (GANs) aimed at reconstructing a clear HR image from a blurry LR input. Their approach involves the design of a multi-class GAN capable of managing both text and face images through a singular generator network.

The PCA-SRGAN [58] approach tackles the challenge of image reconstruction by breaking down an image into its principal components and feeding them gradually to the discriminator, thereby simplifying its learning process. This innovative approach aimed at producing more accurate images with clearer contours and fine textures.

Zhang et al. proposed a supervised pixel-wise GAN network [59]. The Pixel-wise GAN method differs from other GAN-based methods in that it operates on a pixel-by-pixel level rather than using a global approach. This allows the network to generate more realistic and detailed images but also makes the training process more complex and computationally expensive.

ESRGAN [60] is a powerful generic super-resolution network based on GAN models. This model is an enhancement of the SRGAN [61], which had previously been proposed. ESRGAN has been used to achieve impressive results in super-resolution for a variety of image types, including face images.

Addressing the limitations of using a single probability value to characterize whether the generated image is a real face image, Zhang et al. [62] developed SPGAN. This method outputs a discriminative matrix with the same resolution as the input images, resulting in more photo-realistic face images, and marking a significant advancement in the field.

Recent advancements in this domain have witnessed the advent of methods such as HiFaceGAN [15], which further refined the super-resolution process using generative models. These methods represented a significant leap in enhancing the quality and realism of generated face images.

Generally, GAN-based FR techniques use a weighted combination of pixel-wise and adversarial loss terms to make a balance between reconstruction accuracy and precision. Hence, the output reveals either signs of smoothness or artifacts depending on the weighting coefficients. To tackle this problem, some researchers proposed taking advantage of GAN-prior networks.

2.2.3 Diffusion Based BFR

2.2.3.1 Denoising Diffusion Probabilistic Models (DDPM)

Denoising Diffusion Probabilistic Models (DDPM), introduced by Ho et al. in 2020 [63], mark an evolution in diffusion-based generative models, a concept initially explored by Sohl-Dickstein et al. in 2015 [64] and further advanced through the Noise-Conditioned Score Network (NCSN) by Yang & Ermon in 2019 [65].

DDPMs are specialized in probabilistic data generation, particularly excelling in image denoising tasks. They employ a complex transformation process involving both forward and reverse (backward) diffusion steps (see Figure 2.4). This process is crucial to understanding how DDPMs systematically convert noise into structured data and vice versa. The forward process adds noise to the data incrementally, while the reverse process methodically removes this noise to reconstruct the original data. Grasping these two processes is essential to appreciate the functionality and effectiveness of DDPMs in Blind Face Restoration and other generative applications.

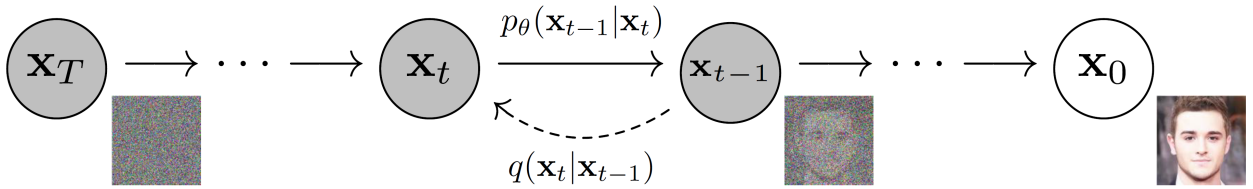


Figure 2.4: Sequential diffusion process in DDPMs: The forward process incrementally adds Gaussian noise to the data, transforming a clean sample into a noisy one. The reverse process methodically denoises the data, reconstructing the original sample from noise. (Image is sourced from [63])

2.2.3.1.1 Forward Process

The forward process in DDPMs represents a noise-adding procedure. Starting with a data point $\mathbf{x}_0$ sampled from the real data distribution $q(\mathbf{x})$, the forward diffusion process gradually introduces Gaussian noise over T steps, resulting in a sequence of increasingly noisy samples $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T$. This process, denoted as $q(\mathbf{x}_{1:T} | \mathbf{x}_0)$, forms a Markov chain:

$$q(\mathbf{x}_{1:T} \mid \mathbf{x}_0) := \prod_{t=1}^T q(\mathbf{x}_t \mid \mathbf{x}_{t-1}). \quad (2.8)$$

At each step in this chain, a conditional Gaussian distribution adds noise to the data. The amount of noise introduced at each step is controlled by a predetermined variance schedule $\beta_1, \beta_2, \dots, \beta_T$:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t} \mathbf{x}_{t-1}, \beta_t \mathbf{I}). \quad (2.9)$$

where $\mathbf{I}$ denotes the identity matrix. As the process progresses and t increases, the data sample $\mathbf{x}_0$ gradually loses its distinguishable features. In the limit, as $T \rightarrow \infty$, $\mathbf{x}_t$ approaches an isotropic Gaussian distribution.

2.2.3.1.2 Reverse Process (Backward Process)

In contrast to the forward process, the reverse process in DDPMs is dedicated to removing noise, effectively reversing the diffusion sequence to transform noisy data back into clean, structured data samples. This reverse process is achieved by using a series of latent variables that form a Markov chain:

$$p_\theta(\mathbf{x}_{0:T}) := p(\mathbf{x}_T) \prod_{t=1}^T p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t), \quad (2.10)$$

where θ denotes the parameters of the model are being learned during the training process. This process begins from a state of pure Gaussian noise, $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. It then progressively constructs less noisy latent variables in a stepwise manner, moving from $\mathbf{x}_{T-1}$ to $\mathbf{x}_{T-2}$, and so on, eventually culminating in the generation of a realistic data sample at $\mathbf{x}_0$. Each step in this reverse sequence follows a Gaussian transition distribution:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)). \quad (2.11)$$

Thus, the reverse diffusion process starts from an initial state of pure Gaussian noise, represented by $\mathbf{x}_T$, to increasingly less noisy states. Over a series of T steps, it methodically refines these latent representations, ultimately restoring the noisy input to a realistic and clean sample at $\mathbf{x}_0$. This carefully calibrated backward diffusion is key to the model's ability to reconstruct HQ, denoised data from noisy inputs.

2.2.3.1.3 Objective Function

The objective function in DDPMs can be represented as a combination of several Kullback-Leibler (KL) divergence and entropy terms. This combination is crucial for modeling the transition between the noisy and clean data (See the detailed process in Appendix B in [64]):

$$\begin{aligned}
L_{VLB} = \mathbb{E}_q[& \underbrace{D_{\text{KL}}(q(\mathbf{x}_T | \mathbf{x}_0) \| p(\mathbf{x}_T))}_{L_T} \\
& + \sum_{t=2}^T \underbrace{D_{\text{KL}}(q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) \| p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_t))}_{L_{t-1}} \\
& - \underbrace{\log p_\theta(\mathbf{x}_0 | \mathbf{x}_1)}_{L_0}].
\end{aligned} \tag{2.12}$$

The KL divergence terms in L_{VLB} (equation 2.12) compare two Gaussian distributions and can be computed in closed form. Given that the forward process variances β_t are fixed to constants, the term L_T is rendered independent of any learnable parameters, thus remaining constant during training and can be disregarded. The term L_0 can be minimized separately as Ho et al. [63] have modeled it using a separate discrete decoder. Consequently, attention is directed towards the term L_{t-1} , for which the posterior conditional distribution is determined using the Bayesian rule:

$$q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) = q(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{x}_0) \frac{q(\mathbf{x}_{t-1} | \mathbf{x}_0)}{q(\mathbf{x}_t | \mathbf{x}_0)}. \tag{2.13}$$

To compute the conditional posterior, we can leverage an elegant property of the forward process, which allows us to directly obtain $\mathbf{x}_t$ through the application of the reparametrization trick. Let $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$, and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Now, DDPM authors, $\mathbf{x}_t$ can be directly obtained as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon_t, \tag{2.14}$$

and $q(\mathbf{x}_t | \mathbf{x}_0)$ is:

$$q(\mathbf{x}_t | \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t) \mathbf{I}). \tag{2.15}$$

By incorporating equation 2.15 at both time t and $t - 1$ into equation 2.13, we can derive the posterior distribution as follows:

$$q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}), \tag{2.16}$$

where the mean and covariance are as follows:

$$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t, \mathbf{x}_0) := \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1 - \bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1 - \bar{\alpha}_{t-1})}{1 - \bar{\alpha}_t}\mathbf{x}_t, \quad \text{and} \quad \tilde{\beta}_t := \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}\beta_t. \quad (2.17)$$

Thanks to the property in equation 2.14, we can represent $\mathbf{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}_t)$ and plug it into the above equation and obtain:

$$\tilde{\boldsymbol{\mu}}_t = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \boldsymbol{\epsilon}_t \right). \quad (2.18)$$

Returning to L_{t-1} term in equation 2.12, our goal is to train a neural network parameterized by θ to minimize the KL divergence between $p_\theta(\mathbf{x}_{t-1} | \mathbf{x}_{t-1})$ and $q(\mathbf{x}_{t-1} | \mathbf{x}_t, \mathbf{x}_0)$. Given that both distributions are Gaussian, this minimization task is equivalent to predicting $\tilde{\boldsymbol{\mu}}_t$ using the neural network. Consequently, the neural network should take $\mathbf{x}_t$ as input and predict $\tilde{\boldsymbol{\mu}}_t = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \boldsymbol{\epsilon}_t)$. Importantly, since $\mathbf{x}_t$ is present both at the input and the output, we can achieve this by reparameterizing the Gaussian noise term to predict $\boldsymbol{\epsilon}_t$ based on the input $\mathbf{x}_t$ at time step t :

$$L_t = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\epsilon}} \left[\frac{(1 - \alpha_t)^2}{2\alpha_t(1 - \bar{\alpha}_t) \|\boldsymbol{\Sigma}_\theta\|_2^2} \|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2 \right]. \quad (2.19)$$

However, Ho et al. [63] observed improved results when employing a simplified objective that disregards the weighting term:

$$\begin{aligned} L_t^{\text{simple}} &= \mathbb{E}_{t \sim [1, T], \mathbf{x}_0, \boldsymbol{\epsilon}_t} [\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)\|^2] \\ &= \mathbb{E}_{t \sim [1, T], \mathbf{x}_0, \boldsymbol{\epsilon}_t} [\|\boldsymbol{\epsilon}_t - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\boldsymbol{\epsilon}_t, t)\|^2]. \end{aligned} \quad (2.20)$$

Thus, the complex upper-bound objective is simplified to a straightforward MSE loss. In this training approach, a single neural network, denoted by $\boldsymbol{\epsilon}_\theta$, is trained to estimate the noise $\boldsymbol{\epsilon}_t$ based on the input $\mathbf{x}_t$ at each time step t .

2.2.3.2 DR2

DR2 (Diffusion-based Robust Degradation Remover) is a cutting-edge diffusion based BFR method that aims to effectively handle complex, real-world image degradations. This approach stands out by first converting degraded images into a basic, degradation-resistant form, then refining them to high-quality images. DR2 operates on the principle that various

degradations ultimately converge to Gaussian noise, allowing for effective semantic capture through iterative denoising steps. The technique showcases robustness against common degradation types.

Figure 2.5 illustrates the DR2 architecture, highlighting the two-stage approach: degradation removal followed by enhancement for final high-quality output. This involves adding noise in the forward direction and progressively removing the noise in the reverse process to eliminate all the degradation, followed by an enhancement module to add a final touch of quality.

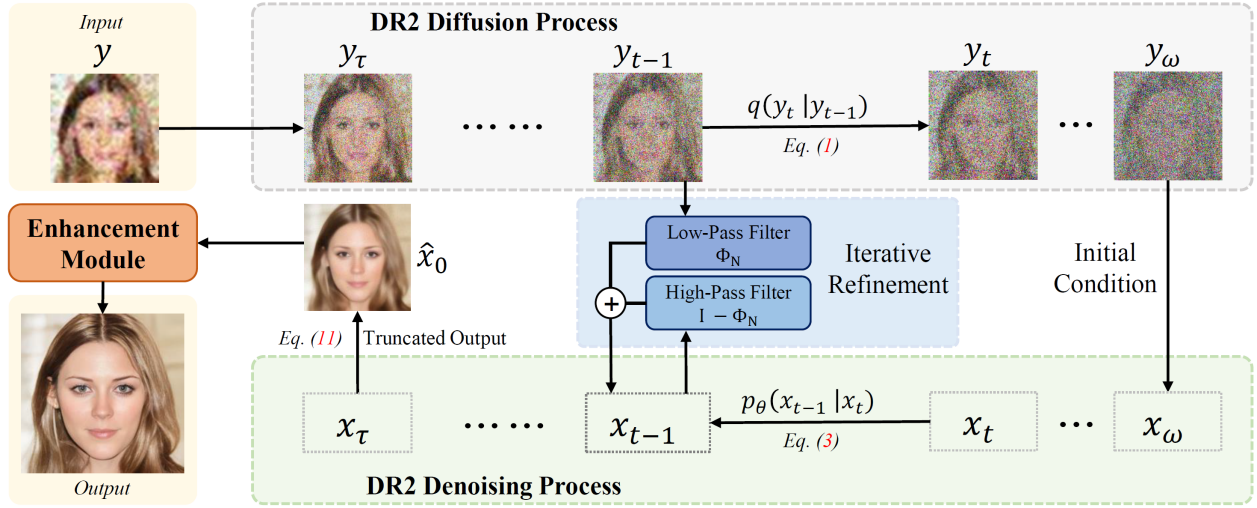


Figure 2.5: Overview of the DR2 method. Image from [1]

Advantages and limitations of DR2 can be summarized as follows:

- **Advantages:**

- **Robust Degradation Handling:** DR2 excels in managing a variety of degradations, translating into consistent performance across diverse degradation scenarios.
- **Quality Preservation:** The two-stage process ensures high-quality restoration generating highly realistic images.
- **Versatility:** DR2's compatibility with various enhancement modules makes it adaptable to a wide range of applications.
- **Innovative Approach:** Leveraging diffusion-based techniques, DR2 introduces a novel perspective in the field of BFR.

- **Limitations:**

- **Complexity:** The intricacy of the DR2 framework can present challenges in terms of computational resources and implementation.
- **Limited generalization power** The effectiveness of DR2 is partly contingent on the performance of underlying pre-trained models, which could be a limiting factor in certain cases.
- **Identity uncertainty** The network may not appropriately preserve the identity of face images during the process

2.3 Prior Based BFR

Prior based BFR methods leverage prior knowledge about facial images to guide the restoration process. These methods utilize different types of priors, including geometric, generative, dictionary, and codebook priors, to enhance the quality and fidelity of the restored faces.

In this section of the thesis, we will delve deeper into each category of prior-based BFR methods and explore the underlying principles, methodologies, advantages, limitations, and performance of each category. By analyzing and comparing these approaches, we aim to gain a comprehensive understanding of the prior-based BFR techniques and their potential applications in restoring degraded facial images.

2.3.1 Geometric Prior Based BFR

Geometric prior-based methods in BFR leverage the unique geometry and spatial distribution of facial features to progressively restore HQ face images. These methods utilize various types of geometric priors, including facial heatmaps [66], facial landmarks [67, 68], facial parsing maps [69], and facial landmarks [67], which are instrumental in reconstructing facial details.

A pioneering work in this field is FSRNet by Chen et al. in 2018 [67], which represents an early attempt to integrate a specific face geometric prior estimation sub-network within a deep learning framework. The process starts with a coarse network for initial HR image recovery, followed by a fine super-resolution network combined with a prior information estimation network. This dual-phase process effectively merges image features with geometric priors for enhanced face restoration.

Following FSRNet, Bulat and Tzimiropoulos [70] introduced Super-FAN, an integrated system for simultaneous facial super-resolution and facial landmark localization. Its core strategy is a joint training approach that enables the network to learn more detailed facial geometric information, primarily through a facial alignment sub-network using heatmap regression.

Yu et al. [66] proposed a method focusing on facial component heatmaps. Their approach, MTUN, features two branches: one for FSR and another for estimating facial component heatmaps, emphasizing the role of facial components like eyes and nose in enhancing face restoration.

Another advancement was made by Kim et al. [68] with their work on progressive FSR via attention to facial landmarks. They developed a lightweight face alignment network to generate landmark heatmaps, which aids in restoring facial details around these landmarks.

Chen et al. [69] further contributed with a progressive model for blind face restoration. They employed semantic-aware style transformation, working in a coarse-to-fine manner for face restoration. This method exemplified the effective use of semantic information in the restoration process.

In a comprehensive approach, Yu et al. [71] utilized multi-geometric priors, combining semantic-level parsing maps, geometric-level facial heatmaps, and reference-level facial dictionaries. This method highlights the efficacy of integrating diverse types of prior information for guiding face restoration.

A significant shift was introduced by Hu et al. [72] with their use of 3D priors for face restoration. Moving beyond traditional 2D priors, their method provides a more detailed parameter description of facial attributes through 3D morphological knowledge, enhancing the overall face restoration performance.

Each of these methods has contributed to the evolution of geometric prior-based BFR, demonstrating a growing complexity and sophistication in the use of facial priors, ranging from landmark and heatmap analysis to the integration of multi-prior learning and 3D morphological insights. However, despite their advancements, geometric prior-based methods have limitations.

One of the main challenges is the dependency on the accuracy and robustness of geometric prior estimation. If the initial estimation of landmarks or heatmaps is inaccurate, especially in severely degraded images, the restoration quality can be significantly compromised. Additionally, these methods may struggle with extreme variations in facial expressions, poses, or occlusions, as the geometric priors might not capture these nuances effectively. Furthermore, the reliance on extensive prior information can make these models computationally intensive,

limiting their practical applicability in resource-constrained environments. Lastly, while 3D priors offer more detail, they also introduce complexity in terms of model training and data requirements, potentially affecting the model’s generalizability to diverse real-world scenarios.

2.3.2 Generative Prior Based BFR

In contrast to the general GAN-based methods that primarily focus on designing various GANs for learning from paired or unpaired data, generative prior-based methods aim to utilize pretrained generative models to enhance BFR performance. The motivation behind this approach is that the pretrained models have already learned useful features and patterns from vast amounts of data, which can be effectively leveraged to improve the restoration quality of face images.

These generative prior-based methods often involve fine-tuning the pretrained generative models on face image datasets to adapt them specifically for BFR tasks. This can lead to better performance in generating HR face images compared to training a GAN from scratch or using other types of general GAN-based methods.

2.3.2.1 Pretrained Generative Models

Pretrained generative models are machine learning models that have been trained on large and diverse datasets before being fine-tuned or adapted for specific generative tasks. These models are a type of transfer learning, where knowledge gained from one task is leveraged to improve performance on another related task. Pretrained generative models have gained significant attention and popularity in recent years, particularly in BFR domain. Some recent examples of generative prior-based methods include the following:

2.3.2.1.1 StyleGAN

StyleGAN [73] is a SOTA generative model known for its ability to generate HQ images. Researchers have explored using StyleGAN as a generative prior for BFR tasks. By fine-tuning StyleGAN on face image datasets, the model can be adapted to generate HR face images from LR inputs. The latent space of StyleGAN provides a rich representation of facial features, enabling the generation of more detailed and realistic restored face images. One of the key components of StyleGAN is its latent space, which provides a rich representation of facial features. The latent space is a multi-dimensional space where each point corresponds to a unique image generated by the StyleGAN model. By navigating and manipulating points in the latent space, it is possible to control the characteristics and appearance of the generated images.

In StyleGAN-based BFR, the goal is to find a point in the latent space that corresponds to a HQ version of a given LQ input image. To achieve this, an optimization process is employed to search for a latent code that, when passed through the fine-tuned StyleGAN generator, produces a HQ face image that closely resembles the LQ input. This optimization process is typically guided by a loss function that measures the similarity between the generated HR image and the LR input.

Once the optimal latent code is found, the fine-tuned StyleGAN generator can produce a HQ face image that is a restored version of the input LQ image. Due to the rich representation of facial features in StyleGAN’s latent space and the model’s ability to generate HQ, high-resolution images, the resulting restored face images are more detailed and realistic than those generated by other BFR methods.

As a core component utilized in this thesis, a thorough analysis of StyleGAN is provided in Appendix [A.1](#).

2.3.2.1.2 BigGAN

BigGAN [74], short for "Big Generative Adversarial Network," is similar to StyleGAN in its ability to generate HQ images. However, BigGAN is designed to generate a wide range of diverse, HR images across various classes, whereas StyleGAN is more focused on generating HQ human faces.

Researchers have explored the use of BigGAN for BFR tasks, taking advantage of its ability to generate diverse, HR images. By fine-tuning BigGAN on face image datasets, the model can be adapted to generate HQ face images with improved fidelity and detail. The fine-tuning process involves training BigGAN on a dataset of HR face images, allowing it to learn the distribution and characteristics of facial features in HR space.

In the context of BFR, the goal is to find a point in BigGAN’s latent space that corresponds to a HR version of a given LR input image. The latent space of BigGAN is a multi-dimensional space where each point corresponds to a unique image generated by the model. By navigating and manipulating points in the latent space, it is possible to control the characteristics and appearance of the generated images.

To achieve BigGAN-based BFR, an optimization process is employed to search for a latent code that, when passed through the fine-tuned BigGAN generator, produces a HR face image that closely resembles the LR input. This optimization process is typically guided by a loss function that measures the similarity between the generated HR image and the LR input.

Once the optimal latent code is found, the fine-tuned BigGAN generator can produce a HR face image that is a restored version of the input LR image. Due to the model’s ability to generate diverse, HR images, the resulting HQ face images have improved fidelity and detail compared to those generated by other BFR methods.

2.3.2.1.3 VQVAE

Vector Quantized Variational Autoencoders (VQVAE) first introduced by [75] is another generative model that has been explored for BFR tasks. VQVAE learns a discrete latent representation of the input data, making it suitable for capturing fine-grained details in face images. Researchers have employed VQVAE as a generative prior for BFR, fine-tuning the model on face image datasets to generate HQ face images from LQ inputs. To understand VQVAE, it is helpful to first consider what traditional Variational Autoencoders (VAEs) do. VAEs are a type of neural network that learn to compress data (e.g., images) into a lower-dimensional space—a process called encoding—and then to reconstruct the data back to its original form—a process called decoding. However, VQVAE introduces a key innovation to this process.

The "Vector Quantized" part of VQVAE refers to how these models handle the encoding step. Instead of mapping input data to a continuous latent space, as traditional VAEs do, VQVAE maps inputs to a finite set of discrete latent vectors or 'codes'. These codes can be conceptualized as fixed anchor points in the latent space. Each of these vectors represents a specific pattern or feature found in the data.

HQ face restoration requires paying attention to intricate details and variations that are often unique to individual faces. VQVAE excels at this because the discrete nature of its latent space allows it to better model and recall these fine-grained details. The discrete codes act like a catalog of facial features that the model can draw from to fill in missing or corrupted parts of a face image.

In practical terms, using VQVAE for BFR involves first training the model on a dataset of face images to develop its discrete latent space. Once trained, the model can take a LQ input image and use its internal 'codebook' of facial features to restore the image, enhancing the resolution, sharpness, and overall fidelity. This makes it particularly suitable for producing HQ reconstructed images from LQ inputs—a task that is critically important in various practical applications, from surveillance to digital entertainment.

Researchers harnessing VQVAE for BFR often fine-tune their models on specific face datasets. This fine-tuning process is about adjusting the pre-trained VQVAE so that its discrete latent vectors are even better at capturing the nuances of human faces. As a result, the model becomes more specialized and thus more proficient at restoring face images, even when they are significantly degraded.

2.3.2.2 GAN Inversion

With the advance of deep generative networks, some innovative GAN models like StyleGAN [73] and StyleGAN V2 [76] have been developed. These GAN models, being trained on huge facial datasets, have demonstrated their capability to produce a vast variety of highly realistic images. Therefore, a growing number of researchers are investigating the potential of using the generative capabilities of these pre-trained models in image reconstruction. GAN-prior networks use the GAN inversion technique to find the latent vector representation of the image, which, when input into a pre-trained GAN, produces the uncorrupted image.

GAN inversion is the task of inverting the input image back to the latent space, where the GAN-prior network can regenerate the original, uncorrupted version of the image. GAN inversion allows for the reconstruction of an image's underlying structure and provides a way to edit or manipulate images in the latent space. The GAN inversion problem can be defined under a unified mathematical formulation. The generator function of a GAN-prior network can be expressed by $\mathbf{G} : \mathbb{R}^d \rightarrow \mathbb{R}^n$ in which the generator G maps a latent vector $z \in \mathbb{R}^d$ to a realistic image $x \in \mathbb{R}^n$. Using a distance metric $l(\cdot)$ in image space or feature space, we can define the GAN inversion problem by the following formula:

$$z^* = \underset{z \in \mathbb{R}^d}{\operatorname{argmin}} \mathcal{L}(\mathbf{G}(z), x). \quad (2.21)$$

There are two main approaches to solving the GAN inversion problem.

1. Optimization-based methods, which aim to directly optimize the latent space to discover the most optimal latent vector that minimizes the distance metric.
2. Learning-based techniques, where an encoder learns the mapping function from the image into the latent space.

2.3.2.3 Optimization-Based GAN-Prior BFR

One way to restore images using GAN prior networks is to use the optimization-based GAN inversion techniques to find the latent code which best can be mapped to the HR image through the GAN prior network, which after down-grading can match the LQ input image. PULSE [9] and mGANprior [77] are two of the most famous FR techniques that utilize optimization-based GAN inversion techniques.

2.3.2.3.1 PULSE

PULSE (Photo Upsampling via Latent Space Exploration) [9] was the first face restoration method that used a pre-trained StyleGAN network to address the challenge of generating HR, realistic images at resolutions that were previously unattainable. Unlike the conventional methods which relied on supervised learning of the relationship between pairs of LR and HR images, The PULSE method conducts the super-resolution task in an entirely self-supervised manner. Instead of using a loss function on HR images to gradually add details to the LR image, they exploited a downscaling loss to guide the optimization process to explore the latent space of a generative model.

Although the PULSE method can be used with any GAN prior network, the authors suggest using the StyleGAN network, due to its rich latent space, which allows for the effective expression of various features, particularly in generating facial images. The algorithm searches the latent space of the pre-trained StyleGAN network to find the optimal latent vector that results in an HR image that downsamples correctly. An HR image is defined to downsample correctly if the downsampled version of the generated HR image matches the input LR image. This solution can be formally expressed as follows:

$$z^* = \underset{z \in \mathcal{Z}}{\operatorname{argmin}} \operatorname{dist}(\mathbf{G}(z) \downarrow_r, l) \quad (2.22)$$

in which z^* is the optimal latent vector in the latent space $\mathcal{Z}$, GAN prior model is denoted by $\mathbf{G}$, and the input LR image and the scaling factor are represented by l and r respectively. The $\operatorname{dist}(\cdot)$ function denotes a dissimilarity distance between the LR image and the downsampled version of the output.

The PULSE algorithm solves equation 2.22 using an iterative optimization process. It starts from a random initial latent vector and updates the latent code based on the direction of the SGD of a loss function. The loss function is defined as the norm 1 between the input LR image and the downsampled version of the StyleGAN output:

$$\operatorname{dist}(\mathbf{G}(z) \downarrow_r, l) = \|\mathbf{G}(z) \downarrow_r - l\|_1 \quad (2.23)$$

This loss function guarantees the solution matches the input LR image. figure 2.6 illustrates this very well. As expected, all StyleGAN outputs are located on the natural image manifolds.

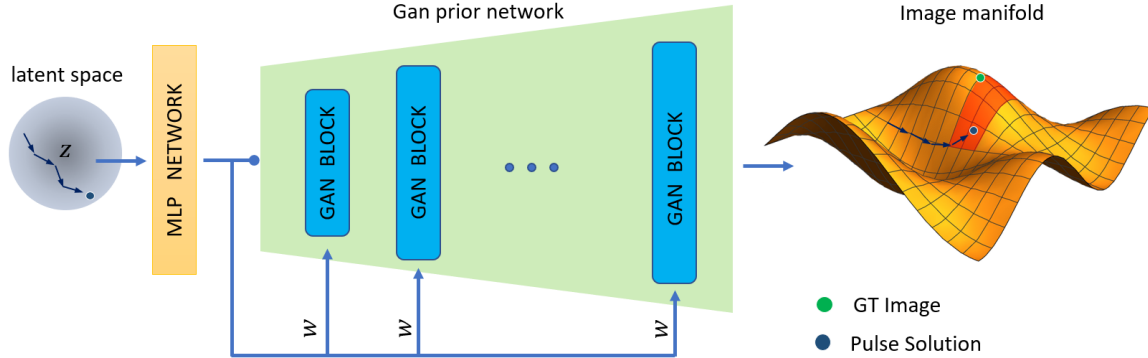


Figure 2.6: The overall structure of PULSE.

However, within the generative model’s manifold used by PULSE, only a small subset corresponds to valid solutions — HR images that accurately downsample to the same LR image. The PULSE method aims to find a solution within this subset, but this solution may still significantly diverge from the actual ground-truth image. This discrepancy is particularly pronounced when the LR image lacks sufficient resolution to retain essential details. Moreover, the iterative nature of inferring an HR image from an LR input in PULSE is both time-consuming and resource-intensive, limiting its utility in real-time applications or on devices with constrained computational resources.

Additionally, the PULSE method, while innovative, does not guarantee the preservation of the original identity in facial reconstruction tasks, as the generated HR images might represent a different but plausible identity that corresponds to the LR input. This raises concerns in applications where fidelity to the original subject is paramount.

The pros and cons of the PULSE method are summarized as follows:

- **Advantages:**

- **Exceptional Detail:** PULSE can generate HR images with exceptional detail, surpassing previous capabilities in image upscaling.
- **Realism and Credibility:** The synthesized images are not only realistic but also perceptually credible, conforming to the natural image manifold structure.

- **Unsupervised Operation:** Operating in an unsupervised manner, PULSE circumvents the need for LR-HR image pairs during training, granting it versatility and strong generalization across varied types of image degradation.

- **Disadvantages:**

- **Computational Demand:** The PULSE method’s inference phase is computationally demanding, posing challenges for time-sensitive applications and devices with limited processing power.
- **Risk of Inaccurate Representation:** There’s a risk of producing HR images that do not faithfully match the ground-truth, especially when starting from extremely LR inputs. This issue is compounded in contexts requiring accurate identity preservation.

2.3.2.3.2 Multi-code GAN prior

The mGANprior (Multi-Code GAN Prior) approach, introduced by Gu et al. [77], represents a significant advancement in leveraging pre-trained GANs as an effective prior for diverse image processing tasks. This method is particularly adept in applications such as FSR and BFR.

Central to mGANprior is the concept of optimizing multiple latent codes to enhance the model’s expressiveness. The innovation lies in decomposition of a latent code into multiple components. Each of these components is responsible for controlling distinct aspects of the generated image. This decomposition allows for more granular control over the image generation process.

By optimizing each of these components, mGANprior facilitates the production of a wider variety of intermediate features. When these features are combined, the model can generate more diverse and expressive outcomes. This capability is a considerable improvement over traditional single-code GAN models, enabling more precise manipulation and refinement of generated images.

The authors of Image2StyleGAN [78] explored the complexities inherent in embedding an image into a single latent vector, or style vector, within a GAN’s latent space. They highlighted that the latent space of a GAN is not only highly complex but also multi-dimensional, where different regions correspond to a wide range of visual features. This complexity poses a challenge in accurately capturing and representing diverse image characteristics within a single vector.

To overcome this challenge, the team behind Image2StyleGAN proposed an innovative approach for embedding images into the StyleGAN latent space. Their method involves optimizing over multiple latent vectors instead of just one. Correspondingly, they introduced a set of style vectors, each designed to modulate different modules of StyleGAN in distinct ways. This method allows for a more nuanced and precise control over how various aspects of an image are represented in the latent space. By leveraging multiple vectors, this approach effectively captures the multifaceted nature of images, enabling a more faithful and detailed representation within StyleGAN’s framework.

Both mGANPrior and Image2StyleGAN methodologies are centered around addressing the GAN inversion challenge, with a specific focus on identifying multiple optimal latent codes. This approach is a departure from traditional methods that typically seek a single latent code representation. The core of their strategy lies in solving an optimization problem, as outlined in Equation 2.24, which aims to accurately map a given image back into the codes in the latent space of a GAN. This process involves finding a set of latent codes that, when passed through the GAN, can reconstruct the original image with high fidelity.

$$z_1^*, z_2^*, \dots, z_{N^*} = \underset{z_1, z_2, \dots, z_N \in \mathcal{Z}}{\operatorname{argmin}} \mathcal{L}(\mathbf{G}(z_1, z_2, \dots, z_N), x). \quad (2.24)$$

where x is the input image, $\mathcal{Z}$ denotes the latent space, and $\mathcal{L}(\cdot)$ denotes the objective function. This objective function is a combination of pixelwise ℓ_2 loss and perceptual loss terms, where the pixelwise loss ensures pixel-level accuracy and the perceptual loss preserves high-level image features such as textures and edges, contributing to the overall realism and fidelity of the reconstructed image. Figure 2.7 depicts the general architecture of optimization-based GAN prior image restoration algorithms, that exploit multiple latent vectors in the restoration process.

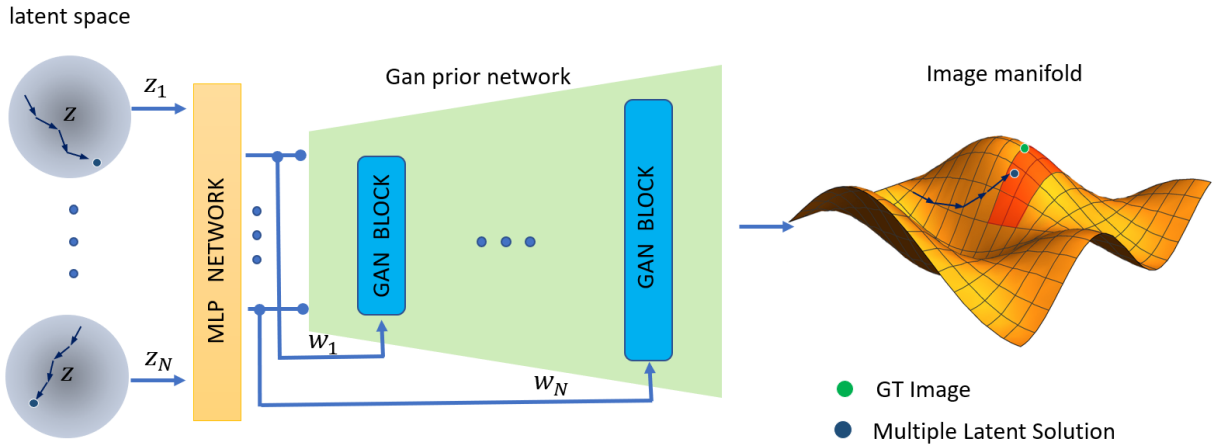


Figure 2.7: Optimization-based GAN priors with multi-Code architecture employed in algorithms such as mGANprior [77] and Image2StyleGAN[78].

The pros and cons of GAN priors with multi-code architecture, are as follows:

- **Advantages:**

- **Enhanced Flexibility:** Compared to PULSE utilizing a single latent code, multiple latent codes offer greater flexibility in image manipulation, allowing for more detailed and nuanced control over the generated images.
- **Improved Quality:** This approach often results in higher-quality image reconstructions, as different aspects of an image can be optimized independently.
- **Broader Applicability:** The multi-code architecture can be applied to a wide range of tasks, including image restoration, super-resolution, and even artistic style transfer.

- **Disadvantages:**

- **Increased Complexity:** Managing and optimizing multiple latent codes can add complexity to the model, potentially making it harder to train and fine-tune.
- **Computational Intensity:** The need to optimize several latent codes simultaneously can lead to increased computational demands, making the method less suitable for real-time applications.
- **Risk of Specificity:** There’s a possibility that using multiple latent codes could lead the model to overly focus on certain image features, potentially leading to less diversity in the generated images or reconstructions that may not comprehensively capture all aspects of the original image.

2.3.2.4 Learning-based GAN Prior BFR

One effective approach in applying GAN prior networks to the BFR task involves training an encoder network to directly embed low-quality input images into the latent space. This method circumvents the need for computationally intensive optimization to determine the latent code during inference, thereby mapping LQ images directly to the latent space. As a result, this approach is well-suited for real-time applications and can be feasibly implemented on devices with limited computational power, such as smartphones. However, it’s important to note that relying solely on learning an optimal mapping from LQ to HQ pairs may not always yield consistent outputs. The efficacy of this system largely depends on the diversity and representativeness of the training data, as well as the accuracy of the degradation model, which could potentially limit its generalization capabilities.

To enhance the robustness of this method, recent research [17, 79, 80] has focused on conditioning the GAN prior network on multi-resolution feature maps generated by the encoder network. This is accomplished through the implementation of skip connections, bridging the encoder and the generative prior network. These connections facilitate the concatenation or manipulation of the GAN blocks' feature maps, allowing for more nuanced image generation. The integration of these multi-resolution feature maps significantly improves the system's performance, enabling it to better accommodate variations in input images and yield more consistent and accurate outputs.

2.3.2.4.1 GPEN

The GPEN (Generative Prior Enhanced Networks) methodology, as proposed by the authors of [17], introduces an innovative approach for BFR using StyleGAN as a GAN prior network. The primary goal of GPEN is to improve the quality of face restoration by harnessing the knowledge from the pre-trained StyleGAN, known for its high-fidelity image synthesis capabilities. Figure 2.8 demonstrates the overall architecture of the GPEN method.

At its core, GPEN integrates the generator network of StyleGAN with a specialized encoder, forming the backbone of the BFR architecture. This system employs a multi-scale feature extraction network, which processes LQ input images and extracts high-level feature representations through convolutional layers. These features are then combined with the intermediate outputs of the StyleGAN generator via skip connections, ultimately producing the restored HR face images.

A crucial advantage of GPEN is its ability to generate HQ face images with fine details and textures, utilizing StyleGAN's robust pre-trained model in conjunction with an effective feature extraction network. To further enhance image quality, the authors suggested fine-tuning the GAN prior network. This process involves initial training on an HQ face image dataset, followed by fine-tuning for the BFR task, allowing the network to adapt more effectively to BFR-specific requirements.

Additionally, GPEN utilizes a multi-scale perceptual adversarial loss during training. This loss function combines adversarial loss with perceptual loss, which assesses differences in feature representations between the generated images and ground-truth HR images. The incorporation of this hybrid loss function ensures that GPEN not only achieves high fidelity in pixel-wise accuracy but also delivers perceptually convincing results that closely resemble the ground-truth images.

Below is a summary of the benefits and drawbacks of the GPEN approach:

- **Advantages:**

- **High-Quality Image Restoration:** GPEN is capable of generating highly realistic images, often surpassing the quality achieved by methods like PULSE, especially in terms of texture and detail in face restoration.
- **Efficient Inference:** The method offers low inference time complexity, making it suitable for real-time applications and practical deployment in various settings.
- **Versatile Degradation Handling:** As a blind restoration approach, GPEN effectively handles a broad spectrum of degradations. It is trained with different blur kernels and varying noise levels, enhancing its robustness and adaptability to diverse image quality challenges.

- **Disadvantages:**

- **Limited generalization power:** GPEN's performance heavily relies on the pre-trained StyleGAN model, which may limit its effectiveness when applied to images that significantly deviate from the training data of StyleGAN.
- **Potential for artifacts:** While GPEN excels in detail and texture, there is a risk of generating minor artifacts or unnatural features.
- **Identity inconsistency:** The identity of the input images may not be perfectly preserved during the process of face hallucination, particularly in challenging restoration scenarios or with very LQ inputs.

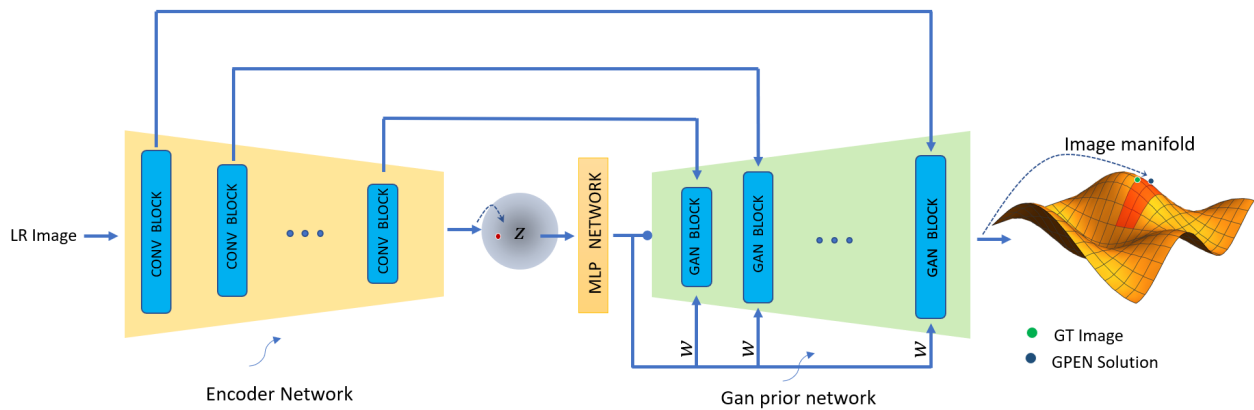


Figure 2.8: The architecture of GPEN method

2.3.2.4.2 GLEAN

The GLEAN method represents a novel approach in the realm of image super-resolution, utilizing StyleGAN as the underlying generative model. Distinctively, GLEAN employs an encoder network that maps LR input images into a series of style codes. These codes are then utilized by the StyleGAN generator to produce HR outputs. This methodology diverges from the traditional use of StyleGAN, which typically relies on a fixed set of style codes. Instead, GLEAN assigns multiple style codes to each LR input, thereby amplifying the expressiveness of the StyleGAN prior network.

In the GLEAN architecture, each block of the StyleGAN generator is fed with a unique style code, enabling the synthesis of more varied textures and details in the generated images. This is a significant advancement over standard approaches, as it allows for a richer and more diverse representation in the super-resolution process.

Furthermore, GLEAN integrates convolutional feature maps extracted from the encoder into the StyleGAN blocks. This integration serves as an additional guide in the reconstruction process, aiding in the accurate depiction of both the overall structure and intricate local details of the images. To facilitate this, each StyleGAN block incorporates a 1×1 convolutional layer, merging the GAN prior features with the encoder feature maps. The amalgamated features are then channeled to a decoder network, which is responsible for reconstructing the final output image.

Figure 2.9 demonstrates the GLEAN architecture, where it's evident that an additional decoder network is employed. This is a strategic deviation from directly generating outputs from StyleGAN. According to the authors, this novel approach enhances the fusion of features from the latent code bank and the encoder, leading to superior image quality in the super-resolution output.

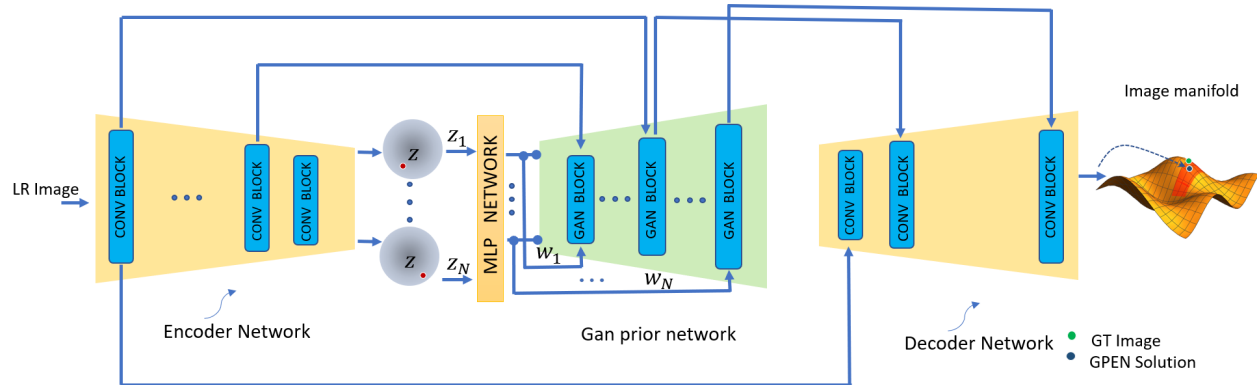


Figure 2.9: Overview of GLEAN method

The benefits and potential drawbacks of the GLEAN method can be expressed below:

- **Advantages:**

- **Enhanced Control and Expressiveness:** GLEAN employs a bank of latent codes for each StyleGAN block, offering enhanced control over the image generation process and improved expressiveness in the final output.
- **High Fidelity and Quality:** The method conditions the generator on convolutional features from the encoder, providing additional guidance on local structures and resulting in high-fidelity and quality results.
- **Flexibility Across Categories:** GLEAN is adaptable to various image categories, thanks to its ability to switch between different GANs, enhancing its versatility.
- **Efficient Inference:** The method offers low inference time complexity, making it suitable for real-time applications and practical deployment in various settings.
- **Efficient Image Upscaling:** Unlike traditional GAN inversion methods requiring extensive optimization, GLEAN achieves large-factor image super-resolution with a single forward pass, streamlining the upscaling process.

- **Disadvantages:**

- **Training Complexity:** The inclusion of an encoder-bank-decoder architecture with multi-resolution skip connections adds complexity to the training process.
- **Limited generalization power:** GLEAN’s effectiveness is contingent on the quality of pre-trained GANs, potentially limiting its performance if the underlying GANs are not robust or diverse enough.
- **Identity uncertainty** The network may not appropriately preserve the identity of face images during the process.

2.3.2.4.3 GFP-GAN

The GFP-GAN [80], or Generative Facial Prior GAN, introduces a unique approach for BFR, aiming to strike a balance between realism and fidelity in a single forward pass. GFP-GAN is comprised of two primary modules: a degradation removal module and a pre-trained face GAN as a facial generative prior. The connection between these modules is facilitated through direct latent code mapping and Channel-Split Spatial Feature Transform (CS-SFT) layers in a coarse-to-fine manner.

The CS-SFT layers in GFP-GAN perform spatial modulation on part of the feature maps, allowing the unmodulated features to pass through directly, which aids in preserving critical information. This method effectively combines the generative prior with high fidelity preservation. The GFP-GAN further employs a facial component loss with local discriminators to enhance perceptual facial details. Additionally, it utilizes an identity-preserving loss to maintain fidelity in the restored faces.

An essential aspect of GFP-GAN is its ability to condition the generator on convolutional features from the encoder. This strategy provides additional guidance for reconstructing both the overall structure and local details of the facial images. Each StyleGAN block utilizes a 1x1 convolutional layer to merge the GAN prior features with encoder feature maps. The combined features then undergo reconstruction in a decoder network to produce the final output.

Figure 2.10 illustrates the GFP-GAN architecture, where an additional decoder network is employed to reconstruct the final image. This approach allows for a more effective fusion of features from the latent bank and encoder, enhancing the overall quality of the restored images.

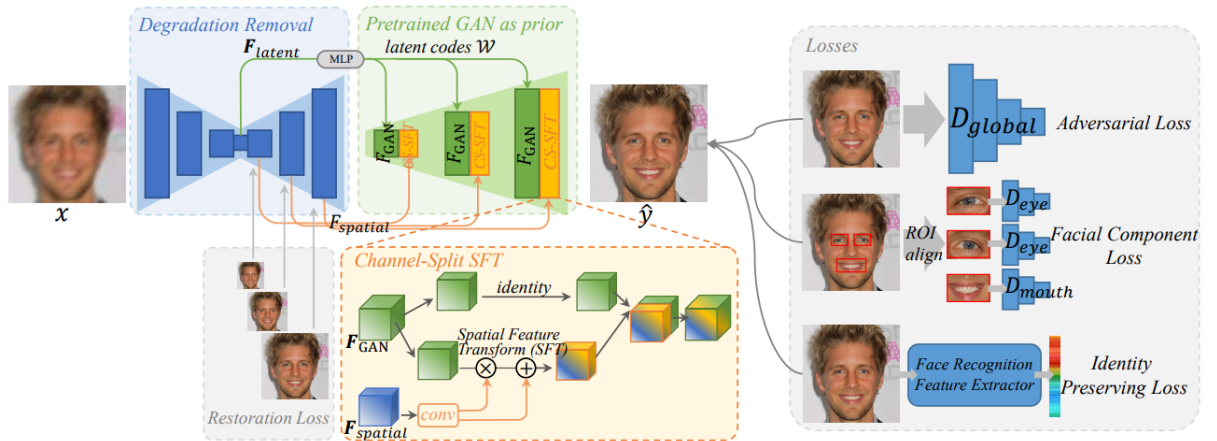


Figure 2.10: Overview of GFP-GAN framework

The GFP-GAN method has been lauded for its ability to restore facial details effectively with a single forward pass, making it a significant contribution to the field of BFR. However, the approach’s reliance on pre-trained GAN models and the complexity of its architecture might pose challenges in specific restoration scenarios. The cons, and pros of GFPGAN are as following:

- **Advantages:**

- **Improved Realism:** GFPGAN is known for generating highly realistic images, particularly useful in fields like art, design, and facial restoration.
- **Versatility in Applications:** Its advanced capabilities make it suitable for various applications, including restoring old or damaged photos and enhancing image quality.
- **Efficient Inference:** The method offers low inference time complexity, making it suitable for real-time applications and practical deployment in various settings.
- **Identity Preservation:** GFPGAN is designed to preserve the identity in face restoration tasks, making it ideal for applications where maintaining the original look is crucial.

- **Disadvantages:**

- **Complexity:** The advanced architecture of GFPGAN can make it complex to implement and train.
- **Limited generalization power:** The performance of GFPGAN can be heavily influenced by the quality and variety of the training data, which may affect its generalizability.
- **Identity uncertainty:** Even with the identity preservation loss employed, the network may not appropriately preserve the identity of face images during the restoration process.
- **Challenges with Severe Degradation:** When the input image is severely degraded, GFPGAN may struggle with retaining HQ image and identity preservation, leading to uncertainty in maintaining the original identity features in the restored image.

2.3.3 Dictionary Prior Based BFR

Dictionary prior-based methods in BFR leverage a collection of HQ face images to form a dictionary of facial components or patches. This dictionary, containing a variety of facial features such as eyes, nose, and mouth under different conditions, serves as a guide in the restoration process.

In 2019, Li et al. [81] introduced a novel approach to reconstructing degraded faces. Their method centers on the selection of similar facial components from various face images. It involves identifying matching facial elements across different images that resemble those in the degraded image, which are then utilized for restoration. This method is particularly notable for its ability to effectively restore facial details from extremely LR images with severe degradation, integrating both low-level and high-level semantic loss functions and a specialized adversarial training scheme.

Li et al. [28] further advanced the field in 2020 with the Deep Face Dictionary Network (DFDNet), which incorporates deep learning into the restoration process. DFDNet employs deep multi-scale component dictionaries, created using K-means clustering on significant face components from HQ images. This method is distinctive for its use of deep learning to match and select the most similar component features from these dictionaries, transferring HQ details to LQ face images. Notably, DFDNet addresses the limitation of requiring a HQ reference image of the same identity by using component dictionaries and a confidence score to adaptively fuse dictionary features, enabling more realistic restoration on real degraded images.

Despite the significant advancements in BFR, the dictionary-based methods share a common limitation in their dependence on HQ images for dictionary creation. This reliance may not fully address the diverse restoration needs of severely degraded images, especially in less represented areas like skin texture or hair. This gap underscores the need for dictionaries that encompass a broader range of facial features and conditions, enhancing the efficacy of BFR in handling varied degradation scenarios.

2.3.4 Codebook Prior Based BFR

While dictionary prior-based methods in blind face restoration employ actual image patches for restoration, codebook prior-based approaches utilize abstract encodings of facial attributes. This paradigm shift is driven by the limitations inherent in facial component dictionaries. As evidenced by works like VQFR [73], RestoreFormer [82], and CodeFormer [18], researchers are increasingly adopting Vector Quantized (VQ) codebooks. These codebooks encompass all crucial facial areas, offering a more comprehensive solution. The concept of VQ codebooks was originally introduced by VQGAN [83] and VQVAE [75] to create a dictionary of representative vectors. These vectors serve as a set of visual elements that can be combined to generate images. In the context of blind face restoration, integrating codebook priors has proven effective in reducing uncertainties imposed by degradation [73].

Typically, these methods employ a quantized autoencoder consisting of an encoder, a vector quantization module, and a decoder, consecutively. Quantization is performed on abstract feature maps generated by the encoder by identifying their nearest neighbors in the reference codebooks. The authors in [73] have shown that quantization leads to degradation removal but entails a trade-off between degradation reduction and fidelity enhancement. Raising the encoder’s downsampling factor results in greater degradation reduction but at the expense of reduced fidelity. VQFR [73] recommended using a downsampling level of 32 for the encoder to strike a balance between degradation reduction and fidelity. Additionally, it proposes employing a parallel decoder to further enhance fidelity. CodeFormer [18] chose a downsampling level of 16 to prioritize fidelity. They also utilized a multi-head transformer to reduce degradation and enhance fidelity further. The transformers mapped the nearest-neighbor codes into more accurate code combinations, resulting in improved accuracy. In contrast to CodeFormer, RestoreFormer (Wang et al. 2022) incorporates a multi-head cross-attention layer to facilitate comprehensive spatial interactions between corrupted queries and HQ key-value pairs.

2.3.4.1 VQFR

Figure 2.11 illustrates the general framework of the VQFR network. It comprises an encoder for abstract feature extraction ($\mathbf{z}^d$), a vector quantization module for nearest-neighbor code quantization ($\mathbf{z}_q^d$), and two parallel decoders—the main branch and the texture branch.

The texture-branch decoder reconstructs HQ textural features ($\mathbf{F}_i^t$) from the quantized features. The resulting HQ features suffer from low-fidelity. To improve their fidelity, they undergo warping using Texture Warping Modules (TWM), guided by input features extracted from degraded image at various spatial levels. Finally, the warped features are input into the main decoder to generate the final restored image.

With the VQ codebook serving as a facial dictionary and the parallel decoder design, VQFR can enhance the quality of facial details while maintaining fidelity during face restoration.

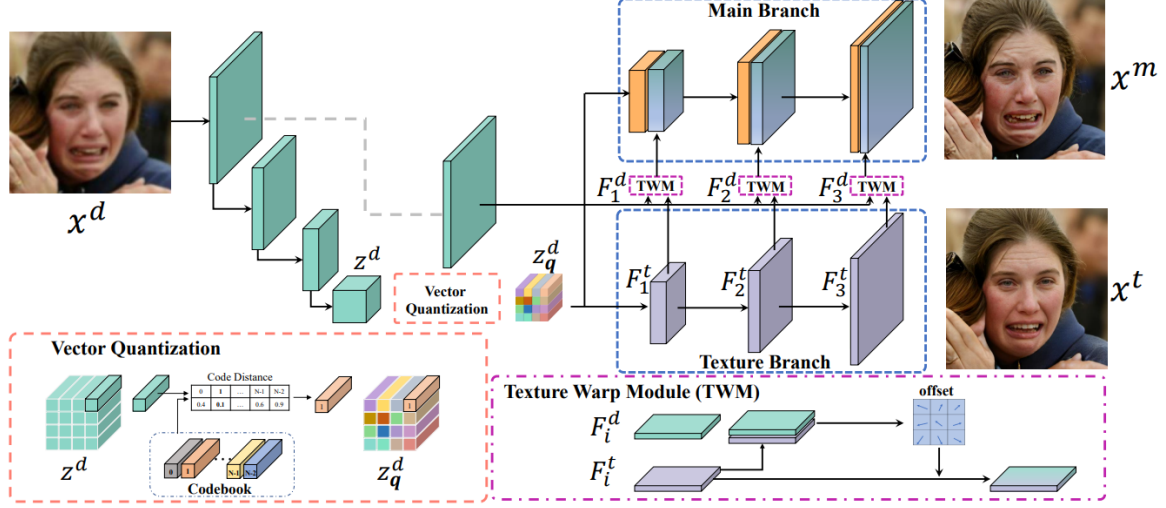


Figure 2.11: The VQFR framework. Image from [73].

2.3.4.2 CodeFormer

CodeFormer employed a pretrained quantized autoencoder initially trained on HQ images. Following this, they performed fine-tuning of the encoder using LQ-HQ image pairs, while keeping the decoder frozen. To enhance the results further, a multi-head transformer was incorporated immediately after fine-tuning the encoder to predict a refined, more quantized code combination.

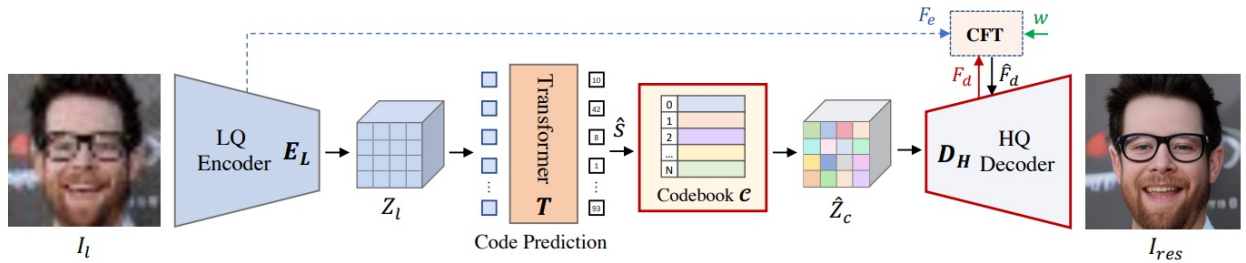


Figure 2.12: CodeFormer framework. Image from [18]

The general framework of the CodeFormer method is illustrated in Figure 2.12. To strike a balance between fidelity and quality, an additional module named the "controllable

feature transformation (CFT) module" was employed. This module regulates the flow of information from the LQ encoder to the HQ decoder. Spatial feature transformation, as referenced in [84], was harnessed within the CFT module. This involved applying a learnable affine transformation based on encoder features to the decoder features, thereby augmenting fidelity.

2.3.4.3 RestorFormer

RestoreFormer [82] adopts a methodology akin to CodeFormer, albeit with a notable distinction. In contrast to CodeFormer's utilization of self-attention mechanisms, RestoreFormer leverages cross-attention mechanisms to amalgamate abstract features both prior to and post quantization. Figure 2.13 illustrates the general framework of RestoreFormer.

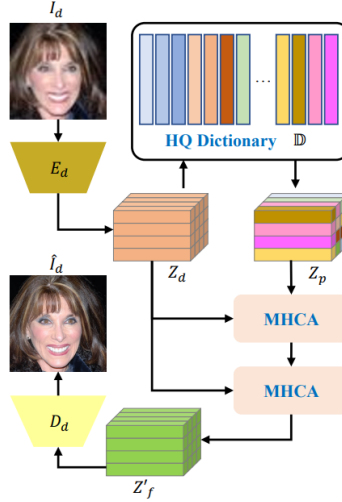


Figure 2.13: RestoreFormer framework. Image from [82].

In a manner reminiscent of CodeFormer, Restoreformer employs a pretrained quantized encoder-decoder architecture. The encoder, designated as E_d , is responsible for encoding the degraded image into abstract features represented as z_p . These abstract features then undergo nearest neighbor quantization, yielding HQ features denoted as z_d . What sets RestoreFormer apart is its utilization of these components in a cross-attention framework.

In this framework, the z_d code assumes the role of the query, while z_p serves as the key and values. This ingenious approach efficiently combines HQ, albeit low-fidelity features obtained after quantization with high-fidelity yet LQ features from before quantization. The result is an accurate amalgamation of code, denoted as z'_f , which is subsequently mapped back to reconstruct the image through the decoder. This distinctive cross-attention mechanism forms the core of RestoreFormer's approach, enabling it to achieve its goal of image restoration.

2.4 Datasets for BFR

In the field of BFR, the choice of dataset plays a crucial role in the development and evaluation of BFR algorithms. HQ datasets provide a solid foundation for training, validating, and testing BFR techniques, ensuring their effectiveness and robustness in real-world applications. A suitable dataset should contain a diverse set of images with varying facial expressions, lighting conditions, and poses to effectively represent the complexities of human faces.

In recent years, many face image datasets have been used for BFR, which differ in various aspects, such as the number of samples, facial characteristics contained, and others. It is worth noting that most datasets primarily provide HQ face images. To use these datasets for training and evaluating BFR models, corresponding LQ face images need to be synthesized using a degradation model (eg., using Equation 1.3).

In this section, we will discuss some of the widely-used datasets for BFR, highlighting their unique characteristics and contributions to the development of BFR techniques, as well as the necessary preprocessing steps involved in adapting them for BFR tasks.

2.4.1 CelebA

The CelebFaces Attributes Dataset (CelebA)[85] is a widely-used large-scale face attributes dataset, consisting of more than 200,000 celebrity images, each with 40 attribute annotations. The images in CelebA exhibit significant pose variations and background clutter, making it a valuable resource for training and evaluating BFR models that can handle complex real-world scenarios.

CelebA offers substantial diversities, large quantities, and rich annotations, which include:

- 10,177 unique identities, providing a diverse representation of facial features, expressions, and appearances.
- 202,599 face images, ensuring a large sample size for more robust and generalizable BFR models.
- 5 landmark locations and 40 binary attributes annotations per image, enabling researchers to explore additional tasks such as facial attribute recognition and facial landmark detection.

Given its comprehensive nature and extensive collection of images, CelebA has become a popular choice for researchers and practitioners working in the domain of BFR.

2.4.2 CelebAMask-HQ

CelebAMask-HQ [86] is a large-scale face image dataset that consists of 30,000 HQ face images selected from the CelebA dataset, following the same procedure as CelebA-HQ. Each image in this dataset is accompanied by a segmentation mask of facial attributes corresponding to CelebA. The masks in CelebAMask-HQ have been manually annotated, with a resolution of 512×512 , and include 19 classes representing various facial components and accessories, such as skin, nose, eyes, eyebrows, ears, mouth, lip, hair, hat, eyeglass, earring, necklace, neck, and cloth.

2.4.3 HELEN

The HELEN dataset [87] is a comprehensive facial image dataset that focuses on facial landmarks and contours. It is composed of 2330 face images of 400×400 pixels with labeled facial components generated through manually-annotated contours along eyes, eyebrows, nose, lips, and jawline. The dataset was specifically designed to improve the performance of facial landmark detection algorithms.

The images in this dataset are diverse, including various lighting conditions, occlusions, expressions, and poses, making it suitable for training and evaluating robust BFR models. In addition to the 2330 HR images, the dataset comes with detailed annotations of facial landmarks, which encompass the eyes, eyebrows, nose, lips, and jawline. These annotations are created by carefully tracing the contours of each facial component manually, ensuring accurate and reliable ground truth data for model training and evaluation.

The HELEN dataset has been widely used in various computer vision tasks, such as facial landmark detection, facial attribute recognition, , and BFR. Its comprehensive annotations and diverse set of images make it a valuable resource for researchers and developers working on facial image analysis and manipulation tasks.

2.4.4 FFHQ

The Flickr-Faces-HQ (FFHQ) dataset [73] is a HQ collection of human face images specifically designed for training and evaluating face-related computer vision tasks. The dataset consists of 70,000 HQ PNG images at a resolution of 1024×1024 pixels, showcasing considerable variation in terms of age, ethnicity, and image background. In addition, it offers good coverage of accessories such as eyeglasses, sunglasses, and hats.

The images in the FFHQ dataset were collected from Flickr, thus inheriting the biases of that website. To create the dataset, the images were automatically aligned and cropped using dlib⁷, and only images under permissive licenses were included. To further refine the dataset,

⁷<http://dlib.net/>

various automatic filters were applied to prune the set, and finally, Amazon Mechanical Turk was utilized to remove occasional statues, paintings, or photos of photos.

The FFHQ dataset has been widely used in various facial analysis and manipulation tasks, such as face generation, and BFR, due to its HQ images and rich diversity in terms of subjects and attributes.

2.4.5 LFW

The Labeled Faces in the Wild (LFW) [88] dataset is a publicly available benchmark for face verification, also known as pair matching. Designed to study the problem of unconstrained face recognition, this dataset contains more than 13,000 face images collected from the internet. Each face is labeled with the name of the person pictured, with 1680 individuals having two or more distinct photos in the dataset. The only constraint applied to these images is that they must be detected by the Viola-Jones face detector [89].

There are currently four different sets of LFW images, including the original set and three types of "aligned" images. The aligned images consist of "funneled images" [90], LFW-a (an unpublished method of alignment), and "deep funneled" images [91]. Among these variations, LFW-a and the deep funneled images yield better results for most face verification algorithms when compared to the original images and the funneled images [90].

The LFW dataset's diversity and unconstrained nature make it a valuable resource for researchers and developers working on face verification and recognition tasks. By providing a challenging benchmark, it helps drive advancements in the field and improve the performance of facial analysis algorithms in real-world applications.

2.4.6 VGGFace

The VGG Face dataset [92] is a comprehensive collection of 2.6 million face images representing 2,622 distinct individuals. Designed to support the development of face recognition technology in the academic research community, it addresses the gap in data availability between academic researchers and large corporations like Google and Facebook.

The dataset primarily consists of images of celebrities, public figures, actors, and politicians, with names extracted from the Internet Movie Database (IMDB) celebrity list, ranked by popularity. This selection criterion ensures a diverse range of individuals with varying facial features, expressions, and appearances, which is crucial for training robust face recognition models.

An important aspect of the VGG Face dataset is its large size, which allows researchers to train deep learning models that can generalize well to real-world scenarios. Moreover, the

dataset covers a wide range of variations in terms of age, ethnicity, and image background, providing a rich source of information for researchers to evaluate their models' performance across different demographics and conditions.

In addition to the image data, the VGG Face dataset includes metadata such as ethnicity, age, and kinship information collected from IMDB. This additional information can be utilized for various auxiliary tasks in face recognition research, such as age estimation or kinship verification, further enriching its value for academic research.

Although the VGG Face dataset serves as a valuable resource for academic research in the field of face recognition technology, it is essential to consider the ethical implications surrounding the use of such data, as the individuals included did not provide consent for the use of their biometric information.

2.4.7 VGGFace2

The VGGFace2 dataset [93], introduced as an enhancement of its predecessor, the VGGFace dataset, offers a wide array of face images tailored for advancing the domain of face recognition. This dataset bridges the chasm between earlier datasets and the pressing need for diverse, real-world reflective data.

Predominantly sourced from Google Image Search, this dataset envelops an impressive volume of images, encompassing a diverse spectrum of individuals. Like its predecessor, the VGGFace2 dataset leans towards images of notable figures, ensuring that researchers are equipped with a myriad of facial features, expressions, and appearances, all indispensable for honing robust face recognition systems.

A noteworthy characteristic of VGGFace2 is its broad scope in terms of variability:

- **Pose Variations:** From frontal captures to profile shots, this dataset ensures models can navigate the intricacies of varying facial angles.
- **Age:** With images spanning across different age groups, VGGFace2 stands resilient to the myriad nuances introduced by age.
- **Illumination:** Emphasizing real-world applicability, it accommodates images from dimly lit environments to brightly illuminated ones.
- **Ethnicity and Profession:** Diversity remains at its core, with a rich mix of ethnic backgrounds and professions, thereby mitigating model biases and ensuring widespread applicability.

The primary intent behind the dataset is to fortify research endeavors, enabling models to thrive across diverse real-world conditions. Supplementing the image data, VGGFace2 also extends metadata, allowing for nuanced research undertakings like age estimation or kinship verification, mirroring its predecessor’s ethos.

However, as with the VGGFace dataset, it’s imperative to weigh the ethical considerations tied to using such data. The individuals featured in VGGFace2, much like VGGFace, haven’t explicitly consented to the usage of their likenesses, emphasizing the need for responsible usage and evaluation in research circles.

The above-mentioned datasets each provide a diverse and large-scale collection of face images with varying characteristics. They differ in the number of images, attributes, landmarks, parsing maps, and identity details, which cater to distinct research requirements in the BFR field. Table 1.1 presents a summary of these public face image datasets commonly used in BFR research.

Table 2.1: Summary of Public Face Image Datasets for BFR.

Dataset	#images	#Attributes	#Landmarks	Parsing maps	Identity
CelebA	202,599	40	5	×	✓
CelebAMask-HQ	30,000	×	×	✓	✓
Helen	2,330	×	194	✓	×
FFHQ	70,000	×	68	×	×
LFW	13,233	73	×	×	✓
VGGFace	2,622,000	×	×	×	✓
VGGFace2	3,310,000	×	×	×	✓

2.5 Image Quality Assessment

Image quality, as it pertains to visual attributes and viewer perception, can be defined in several ways depending on the measurement methods employed. Image quality assessment (IQA) metrics play a crucial role in evaluating the performance of BFR techniques. These metrics help quantify the quality of reconstructed images compared to the original HQ images. The IQA methods can be broadly categorized into subjective methods, which rely on human perception of an image's naturalness and quality, and objective methods, which involve quantitative techniques for calculating image quality numerically [94].

While quality-related visual aspects of an image often provide a good measure, this approach can be resource-intensive, particularly when dealing with large datasets [95]. Consequently, objective methods are more suitable for BFR and computer vision tasks. According to Saad, Bovik, and Charrier [96], IQA methods are primarily divided into three categories: reference image-based, features from the actual image, and blind IQA without any information about the ground truth. In this section, we delve deeper into the IQA methods predominantly used in the domain of BFR.

2.5.1 PSNR

In information systems, the peak signal-to-noise ratio (PSNR) serves as a metric to analyze the signal power against the noise power, predominantly in images. The PSNR provides a quantitative measure of the compression quality of an image. In BFR, the first step to calculate PSNR is determining the Mean Squared Error (MSE) between the pixel values of the original and the reconstructed images. Once the MSE is determined, PSNR is computed as:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\max^2}{\text{MSE}} \right) \quad (2.25)$$

where $\max$ is the maximum possible pixel value of the image. For an 8-bit grayscale image, $\max = 255$. The MSE between the original and reconstructed images is calculated using:

$$\text{MSE} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n (\text{I}_{\text{ref}}(i, j) - \text{I}_{\text{rec}}(i, j))^2 \quad (2.26)$$

where I_{ref} and I_{rec} denote the original (reference) and reconstructed images, respectively. m and n are the dimensions of the images, and the summations are performed over all pixel positions (i, j) .

While higher PSNR values signify better image quality with reduced distortion, lower values indicate a heightened level of noise and degradation. However, it's worth noting that PSNR might not always correlate well with the human perception of image quality, given its emphasis on pixel-wise differences without accounting for the structural or perceptual features of the images [97]. There are scenarios where the PSNR metric can be deceptive, with the reconstructed image not bearing much visual resemblance to the reference [98, 99]. Consequently, in recent years, metrics that better align with human visual perception have gained higher popularity.

2.5.2 SSIM

The Structural Similarity Index (SSIM) is an objective image quality assessment metric introduced by Zhou and Bovik. [30]. This metric evaluates the similarity between two images based on their structural, luminance, and contrast properties. Unlike PSNR, which is centered around pixel-wise differences, SSIM is influenced by the human visual system's perception, making it a more intuitive metric for capturing image quality from a human perspective.

The SSIM value lies between -1 and 1, providing a measure of similarity between a reference image and a distorted or reconstructed image. A perfect score of 1 indicates that the images are identical in terms of structural content, while a score nearing -1 suggests significant dissimilarity. Higher SSIM values generally indicate better image quality and a stronger resemblance between the two images.

The SSIM index calculation is expressed as:

$$\text{SSIM}(x, y) = \frac{(2 \cdot \mu_x \cdot \mu_y + C_1)(2 \cdot \sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (2.27)$$

where x and y represent the two images under comparison. μ_x and μ_y denote their respective mean intensities. Their variances are represented as σ_x^2 and σ_y^2 , and σ_{xy} captures their covariance. The constants C_1 and C_2 are added to stabilize the division operations, ensuring that the denominator doesn't approach zero.

SSIM has gained popularity in recent years for its better correlation with human visual perception, making it more suitable for various image processing tasks, including image compression, restoration, and super-resolution.

2.5.3 MS-SSIM

Multi-Scale Structural Similarity Index (MS-SSIM) [100] is an extension of the SSIM that takes into account image details at different resolutions by considering multiple scales. This enhancement makes MS-SSIM more suitable for assessing the quality of images in tasks like BFR, where preserving fine details is crucial.

One of the limitations of the original SSIM metric is that it does not reliably account for the uneven distribution of image details. To overcome this limitation, MS-SSIM divides the image into multiple windows, first assesses SSIM for each window separately, and then combines these SSIM values to obtain the final MS-SSIM score.

MS-SSIM calculates the SSIM index at various scales by applying a Gaussian pyramid decomposition to the reference and distorted images. It computes the SSIM values at each scale and then combines them using a weighted geometric mean to obtain the final MS-SSIM score. The score ranges from -1 to 1, with higher values indicating better image quality and stronger correlation with human perception.

In comparison to the original SSIM metric, which only considers a single scale and may be affected by uneven image detail distribution, MS-SSIM is better equipped to capture the structural and perceptual similarities between images at different levels of detail. As a result, it provides a more accurate assessment of image quality, especially in cases where the preservation of fine details and textures is of paramount importance.

The MS-SSIM calculation involves the computation of the SSIM index at multiple scales, typically achieved by down-sampling the images using a Gaussian pyramid decomposition. The final MS-SSIM score is obtained by combining these SSIM values using a weighted geometric mean. The MS-SSIM formula can be represented as:

$$\text{MS-SSIM}(\mathbf{I}_{\text{ref}}, \mathbf{I}_{\text{rec}}) = \prod_{i=1}^N [\text{SSIM}(\mathbf{I}_{\text{ref}}^{(i)}, \mathbf{I}_{\text{rec}}^{(i)})]^{\alpha_i} \quad (2.28)$$

where $\mathbf{I}_{\text{ref}}$ and $\mathbf{I}_{\text{rec}}$ are the reference and reconstructed images, respectively, N is the number of scales, $\mathbf{I}_{\text{ref}}^{(i)}$ and $\mathbf{I}_{\text{rec}}^{(i)}$ represent the reference and reconstructed images at scale i , and α_i are the weighting factors that control the relative importance of SSIM values at each scale.

It is important to note that the formula above is a simplified representation of the MS-SSIM calculation. In practice, the computation of MS-SSIM involves calculating the luminance, contrast, and structure terms of the SSIM metric at each scale, and then combining them using a weighted geometric mean with different exponents for each term.

2.5.4 LPIPS

Learned Perceptual Image Patch Similarity (LPIPS) is a perceptual image quality assessment metric developed by researchers at Google Brain [32]. It's designed to measure the perceptual similarity between a reference image and a distorted or reconstructed image. LPIPS is particularly useful in tasks such as image super-resolution, compression, and restoration, where accurately assessing the visual quality of the reconstructed images is essential.

LPIPS uses deep learning and often employs architectures like VGG [101] or AlexNet [102], leveraging pre-trained weights. The neural network extracts feature embeddings from patches of the reference and reconstructed images. The perceptual similarity is then computed from these features. In essence, LPIPS measures the distance in a deep feature space between two images. A smaller LPIPS score indicates that the images are perceptually closer.

Unlike traditional metrics like PSNR or SSIM, which often rely on pixel-wise comparisons, LPIPS aligns better with human visual perception by emphasizing structural and perceptual differences. This makes it particularly apt for applications where the preservation of fine details and textures is important. Empirical evaluations have showcased its closer alignment with human judgment when juxtaposed with PSNR and SSIM.

However, there are considerations to bear in mind. LPIPS, while powerful, isn't as ubiquitous as metrics like PSNR, SSIM, or MS-SSIM. It's primarily in the domain of research or specific applications requiring aligned image quality assessments. Furthermore, as a learned metric, LPIPS could exhibit sensitivity to the dataset on which its underlying network was trained, which might cause variations in assessment across different domains or contexts.

2.5.5 FID

Fréchet Inception Distance (FID) [31], is a metric used to evaluate the quality of generated images, particularly in the context of image restoration, where the similarity between the distributions of the reconstructed and reference images is measured by comparing the deep feature representations of the two sets of images. A lower FID score indicates that the reconstructed images are more similar to the reference images, while a higher score implies a larger discrepancy between the two sets. The better the visual quality, the smaller the FID.

In contrast to traditional metrics like PSNR and SSIM, which focus on pixel-wise differences and local structural similarities, FID assesses the visual quality of face images in a distribution-wise manner. This approach enables FID to capture higher-level perceptual features and more accurately reflect the subjective perception of image quality.

Rooted in the Fréchet distance, which quantifies the disparity between two probability distributions, FID employs the Inception-v3 network — a pre-trained deep convolutional

neural network[103]. This network extracts high-tier feature representations from both original and restored images. Subsequently, the mean and covariance of these feature representations are determined for both image sets. The FID score is then derived as the Fréchet distance between the two corresponding multivariate Gaussian distributions.

FID has become a popular evaluation metric for tasks like image restoration and generation because it captures both the quality and diversity of the generated images. It is generally considered to be more reliable and informative than other metrics, such as PSNR, or SSIM, as it accounts for both the intra-class and inter-class variability in the generated images and better aligns with human visual perception. The mathematical representation of FID is:

$$\text{FID} = \|\mu_{\text{ref}} - \mu_{\text{rec}}\|^2 + \text{Tr}(\Sigma_{\text{ref}} + \Sigma_{\text{rec}} - 2(\Sigma_{\text{ref}}\Sigma_{\text{rec}})^{1/2}) \quad (2.29)$$

where $(\mu_{\text{ref}}, \Sigma_{\text{ref}})$ and $(\mu_{\text{rec}}, \Sigma_{\text{rec}})$ symbolize the mean and covariance of the feature representations for the reference and generated images, respectively. The Euclidean norm is represented as $\|\cdot\|$, and $\text{Tr}(\cdot)$ stands for the trace of a matrix. The square root operation applied to a matrix is computed element-wise.

2.5.6 NIQE

Natural Image Quality Evaluator (NIQE) [104] is a no-reference image quality assessment metric that quantifies the visual quality of an image without requiring a ground truth reference image for comparison. The metric’s methodology revolves around gauging the distance between two multivariate Gaussian models, one fitting the natural image statistics and the other fitting the evaluated image statistics.

Specifically, the fitting of the multivariate Gaussian model is based on quality-aware features derived from the natural scene statistic model. These features capture the statistical properties of natural images, which can help determine whether the evaluated image exhibits characteristics typically found in HQ images.

The NIQE score is a scalar value that represents the distance between the statistical models of the natural images and the evaluated image. A lower NIQE score indicates that the evaluated image has a higher visual quality and is more similar to the natural image statistics, while a higher score implies a larger discrepancy between the two sets of image statistics, indicating a lower visual quality.

In image processing applications, especially when a reference image is either inaccessible or impractical for use, e.g., real-world image restoration, no-reference metrics like NIQE become indispensable, offering a pragmatic solution to quantify image quality.

2.5.7 MOS

Mean Opinion Score (MOS) is a subjective image quality assessment metric that evaluates the perceptual quality of images through human judgment, offering valuable insights into how people perceive image quality. In contrast to the above objective quantitative metrics, MOS relies on the opinions of quality testers who grade images based on specific criteria, such as sharpness, natural look, and color. The final MOS is obtained by calculating the arithmetic mean of the assigned scores.

Participants in a MOS study typically rate images on a scale, often from 1 to 5, where 1 denotes the poorest quality and 5 the highest. These assessments encompass a variety of criteria such as sharpness, color fidelity, and overall naturalness. The culminating MOS is the arithmetic average of these scores.

In the context of image restoration research, some methods that achieved good objective quality scores under-performed in subjective assessments, particularly in human face reconstruction [17, 105, 106]. As a result, researchers have employed MOS in various studies [107, 108, 109, 17] to measure human perception of image quality.

MOS effectively captures human perception, but it is not without limitations. These include non-linearity between scores, variation in results due to changes in test criteria, and potential human error. The reliability of MOS as an assessment metric is contingent upon the number of human raters involved. A small number of raters can introduce bias into the MOS, whereas a larger pool of raters increases the metric’s reliability and reduces potential biases. Consequently, researchers should consider the number of raters when using MOS to ensure a more accurate representation of human perception.

Chapter 3

Methodology

3.1 Motivation

Let Y be the space of the HQ face images and X be the space of the degraded LQ face images. The aim of BFR is to find the corresponding HQ face image $y \in Y$ of an LQ face image $x \in X$. To achieve this goal, most DNN-based methods learn a function $\Phi(x) \rightarrow y$. However, being a one-to-many inverse problem, multiple potential face images (e.g., $y_1, y_2, \dots, y_n$) in Y can be mapped to from an input x . Our objective is to generate an HQ image that genuinely represents the original LQ input image.

Theoretically, this can be accomplished by having a StyleGAN network [73] pre-trained on huge face datasets and using GAN inversion techniques. Specifically, the latent vector corresponding to the input LQ image can be found by adopting the GAN inversion technique. Then, the corresponding HQ image can be retrieved by feeding the latent vector to the StyleGAN network. Let us denote the latent vector as $\mathbf{z} \in \mathbb{R}^d$, the HQ image as $\mathbf{h} \in \mathbb{R}^n$, the LQ image as $\mathbf{l} \in \mathbb{R}^p$ and the StyleGAN network with $\mathbf{G} : \mathbb{R}^d \rightarrow \mathbb{R}^n$. Then the GAN inversion technique can be formally expressed as follows:

$$\mathbf{z}^* = \underset{\mathbf{z} \in \mathbb{R}^d}{\operatorname{argmin}} \mathcal{L}(\Phi(\mathbf{G}(\mathbf{z})), \mathbf{l}) \quad (3.1)$$

where $\mathcal{L}$ is generally a pixel-wise loss function like MSE and MAE and Φ is the assumed degradation function.

Optimization-based GAN inversion methods require a lot of time and computational resources, which limits their practical applications. Therefore, we adopt the learning-based approach for GAN inversion as the alternative.

For the learning-based GAN inversion, we need a CNN network to learn the latent representation of images. The CNN network denoted by $C : \mathbb{R}^p \rightarrow \mathbb{R}^d$ can directly learn the mapping function from the LQ image $\mathbf{l}$ to the corresponding optimum $\mathbf{z}^*$:

$$\mathbf{z}^* = C(\mathbf{l}). \quad (3.2)$$

We take advantage of an *AutoEncoder* network for the learnable CNN function $C : \mathbb{R}^p \rightarrow \mathbb{R}^d$. The AutoEncoder framework is composed of Encoder and Decoder networks. The Encoder network maps the corrupted image to the low-dimensional latent space, while the Decoder network maps the latent representation back to the image space. Therefore, the Encoder network conducts the GAN inversion task by providing the latent representation of the image, while the Decoder network can remove the uncorrelated noise and artifacts and provides a coarse super-resolved image.

That being said, utilizing only the latent vector generated by the AutoEncoder network might not result in high fidelity. It tends to produce HQ images maintaining the pose, yet compromises on preserving the facial structure, leading to a loss of identity. To mitigate such issues, we introduced the following adjustments in our proposed architecture illustrated in figures 3.2 and 3.3.

3.1.1 Style-Identity Integration

Styles are the building blocks in StyleGAN for facial image generation. The richness and accuracy of these styles are crucial for creating faithful representations. These styles inherently encapsulate essential details such as pose, color, and texture. Augmenting these styles with identity information could significantly improve the generation process. A concept analogous to identity in this context is the *identity vector*, produced by a face recognition network, e.g., ArcFace [33]. This vector is designed to minimize distances between vectors of the same individual while maximizing those between different individuals, effectively capturing unique identity features.

Our approach proposes embedding this identity vector into the styles of StyleGAN. This integration is a novel step towards ensuring that synthetic images not only maintain high visual quality but also stay true to the unique characteristics of each individual’s identity. To perform this integration, we utilize the Style-to-ID Network to infer the identity vector corresponding to a StyleGAN2 style vector $\mathbf{s}$. This process is expressed as $\mathbf{i}_s = F_{\theta}(\mathbf{s})$, where $\mathbf{i}_s$ represents the inferred identity vector and F_{θ} denotes the Style-to-ID function parameterized by θ . Our objective is to minimize the ℓ_2 loss between these inferred identity vectors and the corresponding ground-truth identity vectors ($\mathbf{i}_g$), which are obtained directly from the ground-truth images using the ArcFace network. Given that our image distribution spans c distinct classes of identities, we propose the theorem 1 to underpin our methodology:

Assumption 1. *The ArcFace network generates identity vectors with considerably lower distances⁸ between identity vectors within the same identity class (intra-class distance) as compared to those between identity vectors across different identity classes (inter-class distance).*

Lemma 1. *Let C and I be random variables, and let $f(c, i)$ be a function defined on their joint domain. Under suitable regularity conditions, the following equality holds:*

$$\mathbb{E}_{c \sim P_C, i \sim P_{I|C}} [f(c, i)] = \mathbb{E}_{c \sim P_C, i \sim P_{I|C}, c' \sim P_{C|I}} [f(c', i)].$$

Proof. We begin by writing the left-hand side of the equality in integral form:

$$\begin{aligned} \mathbb{E}_{c \sim P_C, i \sim P_{I|C}} [f(c, i)] &= \int_c P(C = c) \int_i P(I = i | C = c) f(i, c) di dc \\ &= \int_c \int_i P(C = c, I = i) f(c, i) di dc \\ &= \int_c \int_i P(C = c, I = i) f(c, i) \int_{c'} P(C = c' | I = i) dc' di dc \\ &= \int_c P_C(c) \int_i P(I = i | C = c) \int_{c'} P(C = c' | I = i) f(c', i) dc' di dc \\ &= \mathbb{E}_{c \sim P_C, i \sim P_{I|C}, c' \sim P_{C|I}} [f(c', i)]. \end{aligned}$$

□

Theorem 1. *Minimizing the ℓ_2 loss between the style-inferred identity vector $\mathbf{i}_{s_c}$ and the corresponding ground-truth identity vector $\mathbf{i}_{g_c}$ maximizes the variational lower bound of the mutual information between the style vectors and the identity class of images.*

⁸Here, "distance" specifically refers to cosine distance; however, for normalized identity vectors, this effectively translates to Euclidean distance.

For clarity and ease of reference throughout the proof, we provide a list of the symbols used, along with their corresponding descriptions, in the table below:

Table 3.1: Symbols used in the proof and their corresponding descriptions.

Symbol	Description
N	Total number of input images.
c	Identity class of an image.
N_c	Number of images belonging to the identity class c
$\mathbf{l}_c$	Low-quality input image belonging to the identity class c .
$\mathbf{g}_c$	Ground-truth image belonging to the identity class c .
$\mathbf{i}_{g_c}$	Identity vector extracted from $\mathbf{g}_c$, i.e., $\mathbf{i}_{g_c} = \text{ArcFace}(\mathbf{g}_c)$.
$\mathbf{s}_c$	Style vector extracted from $\mathbf{l}_c$.
$\mathbf{i}_{s_c}$	Style-inferred identity vector extracted from style vector, i.e., $\mathbf{i}_{s_c} = F(\mathbf{s}_c)$.

Proof. Our investigation begins with an analysis of the mutual information between identity class c and the style-inferred identity $\mathbf{i}_{s_c}$, which can be expressed as:

$$\mathcal{I}(c, \mathbf{i}_{s_c}) = \mathcal{H}(c) - \mathcal{H}(c \mid \mathbf{i}_{s_c}).$$

Since the entropy $\mathcal{H}(c)$ is constant, maximizing the mutual information $\mathcal{I}(c, \mathbf{i}_s)$ corresponds to minimizing the conditional entropy $\mathcal{H}(c \mid \mathbf{i}_s)$. This conditional entropy can be expressed as:

$$\mathcal{H}(c \mid \mathbf{i}_{s_c}) = \mathbb{E}_{c \sim P(c), \mathbf{i}_{s_c} \sim F(\mathbf{s}_c)} [\mathbb{E}_{c' \sim P(c \mid \mathbf{i}_{s_c})} [-\log P(c' \mid \mathbf{i}_{s_c})]].$$

Given the unavailability of the true posterior $P(c \mid \mathbf{i}_{s_c})$, we approximate it with an auxiliary distribution, denoted as $\mathcal{P}(c \mid \mathbf{i}_{s_c})$:

$$\begin{aligned} \mathcal{H}(c \mid \mathbf{i}_{s_c}) &= \mathbb{E}_{c \sim P(c), \mathbf{i}_{s_c} \sim F(\mathbf{s}_c)} \left[\underbrace{-\text{D}_{\text{KL}}(P(\cdot \mid \mathbf{i}_{s_c}) \parallel \mathcal{P}(\cdot \mid \mathbf{i}_{s_c}))}_{\leq 0} + \mathbb{E}_{c' \sim P(c \mid \mathbf{i}_{s_c})} [-\log \mathcal{P}(c' \mid \mathbf{i}_{s_c})] \right] \\ &\leq \underbrace{\mathbb{E}_{c \sim P(c), \mathbf{i}_{s_c} \sim F(\mathbf{s}_c), c' \sim P(c \mid \mathbf{i}_{s_c})} [-\log \mathcal{P}(c' \mid \mathbf{i}_{s_c})]}_{\mathcal{H}_{\text{ub}}}. \end{aligned}$$

Since the Kullback-Leibler divergence (D_{KL}) is always non-negative, the auxiliary

conditional entropy provides an upper bound for the true conditional entropy. We denote this upper bound by $\mathcal{H}_{\text{ub}}$, which can be further simplified using Lemma 1:

$$\mathcal{H}_{\text{ub}} = \mathbb{E}_{c \sim P(c), \mathbf{i}_{sc} \sim F(\mathbf{s}_c)} [-\log \mathcal{P}(c | \mathbf{i}_{sc})]. \quad (3.3)$$

To derive the auxiliary posterior $\mathcal{P}(c | \mathbf{i}_{sc})$, Bayes' rule is applied to the marginal distribution $\mathcal{P}(\mathbf{i}_{sc})$ and the conditional distribution $\mathcal{P}(\mathbf{i}_{sc} | c)$:

$$\mathcal{P}(c | \mathbf{i}_{sc}) = \frac{\mathcal{P}(c) \mathcal{P}(\mathbf{i}_{sc} | c)}{\mathcal{P}(\mathbf{i}_{sc})}, \quad (3.4)$$

where $\mathcal{P}(c)$ denotes the prior of the identity class, which can be calculated using the following formula:

$$\mathcal{P}(c) = \frac{N_c}{N}. \quad (3.5)$$

We employ the Parzen window estimation method [Parzen, 1962] to approximate the marginal and conditional distributions. By utilizing the ground-truth identity vectors $\mathbf{i}_1, \mathbf{i}_2, \dots, \mathbf{i}_k, \dots, \mathbf{i}_N$, which are derived from feeding all the N ground-truth training images through ArcFace, and applying Parzen window estimation with a Gaussian kernel parameterized by σ^2 , we obtain:

$$\mathcal{P}(\mathbf{i}_{sc}) = \frac{1}{(2\pi)^{d/2} \sigma^d N} \sum_{k=1}^N \exp \left(-\frac{1}{2\sigma^2} \|\mathbf{i}_{sc} - \mathbf{i}_k\|_2^2 \right). \quad (3.6)$$

Similarly, let us define the set $\mathbb{K}_c = \{k | \mathbf{g}^{(k)} \in c\}$, which includes the indices k of all training images belonging to the identity class c with $\mathbf{g}^{(k)}$ denoting the k^{th} ground-truth image in the dataset. Now we can write:

$$\mathcal{P}(\mathbf{i}_{sc} | c) = \frac{1}{(2\pi)^{d/2} \sigma^d N_c} \sum_{k \in \mathbb{K}_c} \exp \left(-\frac{1}{2\sigma^2} \|\mathbf{i}_{sc} - \mathbf{i}_k\|_2^2 \right). \quad (3.7)$$

Putting Eq. 3.5, Eq. 3.6, and Eq. 3.7 inside Eq. 3.4, we obtain the auxiliary posterior distribution as follows:

$$\mathcal{P}(c|\mathbf{i}_{s_c}) = \frac{\sum_{k=1}^N \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{i}_{s_c} - \mathbf{i}_k\|_2^2\right)}{\sum_{k \in \mathbb{K}_c} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{i}_{s_c} - \mathbf{i}_k\|_2^2\right)}. \quad (3.8)$$

Following this, the upper bound $\mathcal{H}_{\text{ub}}$ is derived by substituting the posterior distribution obtained in Eq. 3.8 into Eq. 3.3:

$$\mathcal{H}_{\text{ub}} = \mathbb{E}_{c \sim P(c), \mathbf{i}_{s_c} \sim F(s_c)} \left[-\log \frac{\sum_{k=1}^N \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{i}_{s_c} - \mathbf{i}_k\|_2^2\right)}{\sum_{k \in \mathbb{K}_c} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{i}_{s_c} - \mathbf{i}_k\|_2^2\right)} \right]. \quad (3.9)$$

Subsequently, we use the following triangle inequality for ground-truth identity vectors, $\mathbf{i}_{g_c}$ and $\mathbf{i}_{s_c}$:

$$\forall k, \underbrace{\|\mathbf{i}_k - \mathbf{i}_{g_c}\|_2^2 - \|\mathbf{i}_{s_c} - \mathbf{i}_{g_c}\|_2^2}_{\text{LHS}} \leq \|\mathbf{i}_{s_c} - \mathbf{i}_k\|_2^2 \leq \underbrace{\|\mathbf{i}_{s_c} - \mathbf{i}_{g_c}\|_2^2 + \|\mathbf{i}_{g_c} - \mathbf{i}_k\|_2^2}_{\text{RHS}}. \quad (3.10)$$

We apply the right-hand side (RHS) of the inequality 3.10 to the numerator and the left-hand side (LHS) to the denominator of Eq. 3.9 to obtain a second-level upper bound denoted by $\mathcal{H}_{\text{ub}2}$:

$$\mathcal{H}_{\text{ub}} \leq \underbrace{\mathbb{E}_{c \sim P(c), \mathbf{i}_{s_c} \sim F(s_c)} \left[-\log \frac{\sum_{k=1}^N \exp\left(-\frac{1}{2\sigma^2}(\|\mathbf{i}_{s_c} - \mathbf{i}_{g_c}\|_2^2 + \|\mathbf{i}_{g_c} - \mathbf{i}_k\|_2^2)\right)}{\sum_{k \in \mathbb{K}_c} \exp\left(-\frac{1}{2\sigma^2}(-\|\mathbf{i}_{s_c} - \mathbf{i}_{g_c}\|_2^2 + \|\mathbf{i}_k - \mathbf{i}_{g_c}\|_2^2)\right)} \right]}_{\mathcal{H}_{\text{ub}2}}.$$

which, after extracting the terms involving $\mathbf{i}_{s_c}$, simplifies to the following:

$$\begin{aligned} \mathcal{H}_{\text{ub}2} &= \mathbb{E}_{c \sim P(c), \mathbf{i}_{s_c} \sim F(s_c)} \left[-\log \left(\exp\left(-\frac{1}{\sigma^2}\|\mathbf{i}_{s_c} - \mathbf{i}_{g_c}\|_2^2\right) + K \right) \right] \\ &\equiv \mathbb{E}_{c \sim P(c), \mathbf{i}_{s_c} \sim F(s_c)} [\|\mathbf{i}_{s_c} - \mathbf{i}_{g_c}\|_2^2] \end{aligned} \quad (3.11)$$

where K is a constant independent of $\mathbf{i}_{s_c}$. This equation shows that the ℓ_2 loss results in minimizing the second-level upper bound of the whole conditional entropy variational upper bound.

In the end, we exploit Assumption 1 to complete the proof. Assuming that the intra-class distance between identity vectors is much smaller than the inter-class distance, RHS and LHS of Eq. 3.10 closely approximate the middle term, thereby rendering the second upper bound sufficiently tight. This makes $\mathcal{H}_{\text{ub}_2}$ represent the value of $\mathcal{H}_{\text{ub}}$ rather than merely serving as its upper bound.

Finally, based on the Data Processing Inequality theorem (DPI) [110], we understand that to maximize $\mathcal{I}(c, \mathbf{i}_{sc})$, it is essential to also maximize $\mathcal{I}(c, \mathbf{s})$. This understanding is rooted in the DPI theorem, which indicates that the transformation from $\mathbf{s}$ to $\mathbf{i}_{sc}$ cannot introduce new information about c . Therefore, to ensure that the mutual information between c and $\mathbf{i}_{sc}$ is maximized at the end, the preceding step must be optimally configured to preserve the maximum possible information about c . $\square$

This approach promotes the inclusion of identity information into the style vector $\mathbf{s}$ by shifting the style space to include more of the identity space where the $\mathbf{i}_g$ belongs (see Figure 3.1).

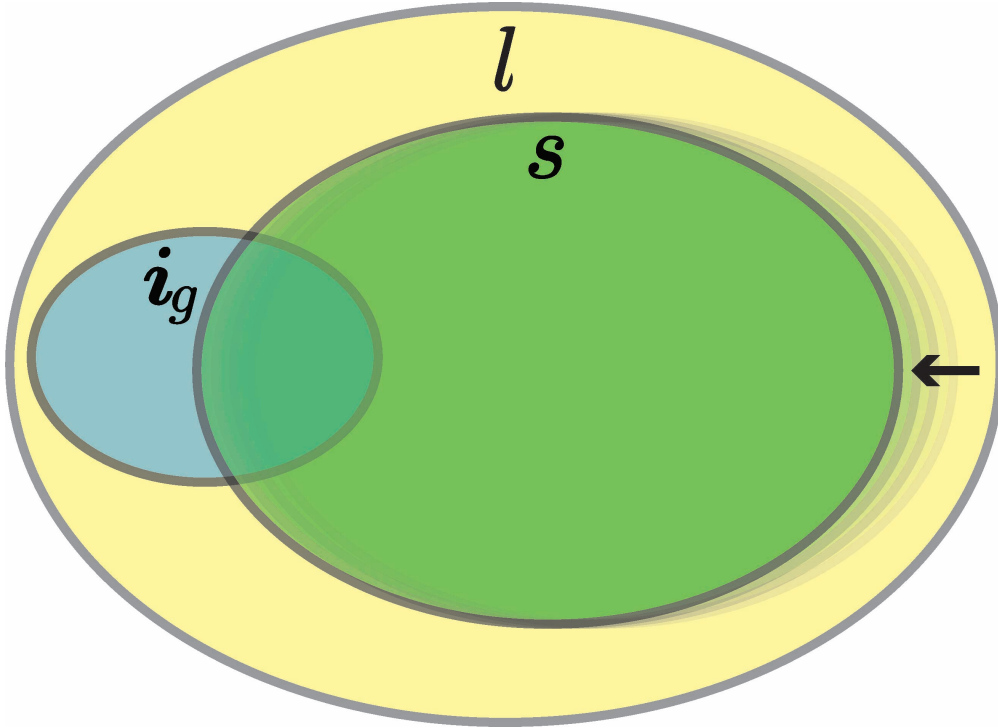


Figure 3.1: Maximizing mutual information for enhanced identity preservation in Style Vectors. The imposition of identity-in-style loss in our methodology facilitates the maximization of mutual information between the identity ($\mathbf{i}_g$) and style ($\mathbf{s}$) vectors. This approach propels the style vector to incorporate and preserve a higher degree of identity information, resulting in a more nuanced and identity-preserving style vectors.

3.1.2 Multi-resolution Feature maps Fusion

StyleGAN2 feature maps, while exhibiting exceptional quality, can potentially compromise fidelity, leading to HQ yet unfaithful outcomes. To mitigate this, we leverage an autoencoder to extract semantic feature maps ($\hat{\mathbf{f}}$) at varying resolutions. Integration of these features in the StyleGAN blocks serves as guidance to create images where face components are better preserved and aligned to that of the original subject.

The decoder network aspires to construct a coarse reconstructed image that generates high fidelity representations where facial structures are maintained. However, due to the ℓ_1 loss deployed as the objective function, the images produced tend to be blurry and not of high quality. Despite this, the high fidelity of these images provides a unique advantage. By integrating the semantic multi-resolution feature maps generated by the decoder with corresponding feature maps derived from the style vectors—which possess high quality but lower fidelity—results in the generation of images where local feature components are better preserved, and the identity of the person in the input image is more accurately retained.

3.1.3 Codebook Prior Image Restorer

To elevate overall image quality and enhance robustness against noise, we integrated the preceding model termed GPR with a codebook-prior-based image restoration framework denoted as CPR. This integration capitalizes on the continuous latent space offered by the GPR framework, facilitating the generation of high-fidelity face images that retain their identity. Simultaneously, the discrete latent space of the codebook-prior model within the CPR framework assures the rendered output image is of exceptional quality.

3.1.4 FTM Module

Considering the potential information loss, particularly regarding identity or fidelity, during the mapping process to the discrete latent space of the codebook-based autoencoder, we introduced the Feature Transformer Module (FTM).

FTM bridges the encoder and decoder at various layers, ensuring that the fidelity and identity of the input remain intact. The primary objective is to mitigate any existing artifacts, culminating in a notable enhancement in image quality.

3.2 Model Architecture

The proposed IGCP model (Identity-embedded GAN and Codebook Priors) adopts a two-stage approach to BFR, uniquely utilizing the continuous latent space of GAN prior networks for identity information fusion and realistic texture synthesis, and the discrete space

of codebook prior networks for producing high-quality output images. The architecture is illustrated in Figure 3.2.

In the first stage, our approach employs a Generative Prior Restorer (GPR) network for initial image restoration. Distinct from existing methods that apply identity-based losses to the final restored image, our work innovates by embedding identity information directly into the style vectors of the StyleGAN [76] network during the generation process. This is achieved through the introduction of an *identity-in-style* loss. GPR synergizes an AutoEncoder (AE) with our novel Identity-Embedded StyleGAN (IES). The AE network supplies the IES with style vectors and spatial feature maps, while an auxiliary Style-to-ID network aids in generation of *identity-infused* style vectors. The identity-infused style vector $\mathbf{s}$, along with multi-resolution feature maps from the AE $\hat{\mathbf{f}}$, are then fed into the StyleGAN generative blocks, where $\mathbf{s}$ serves as a basis to build upon and generate output images, while $\hat{\mathbf{f}}$ steers the generation process to ensure that face components are better aligned to that of the original subject, by merging with the generative features at various resolutions.

Proceeding to the second stage, the approach utilizes a two-component framework known as the Codebook Prior Restorer (CPR) network. This framework comprises a Vector Quantized AutoEncoder (VQAE) [83] for artifact mitigation and to add a final touch of quality, complemented by introducing a Feature Transfer Module (FTM). This module has been proven essential for ensuring fidelity and identity preservation.

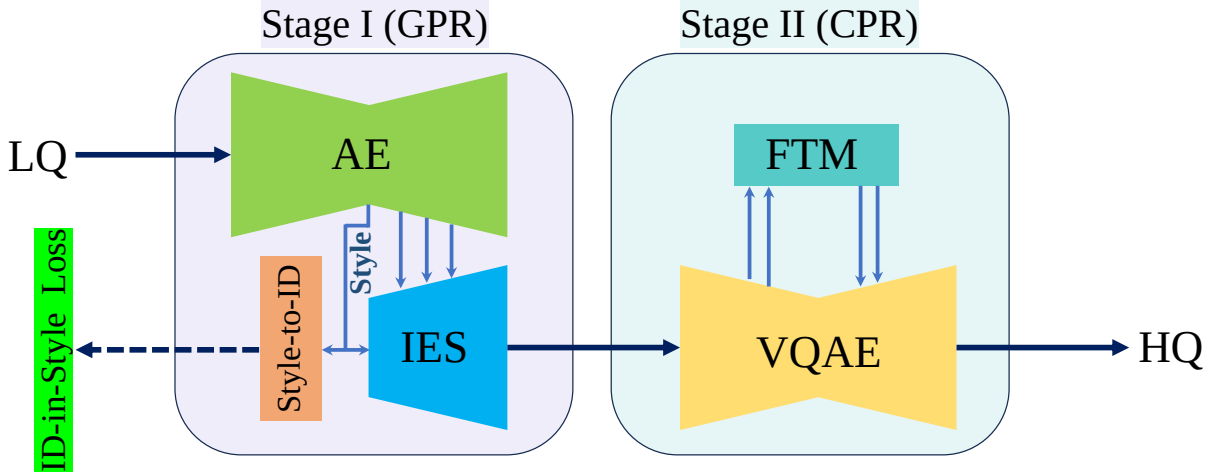


Figure 3.2: IGCP framework overview.

3.2.1 GAN Prior Restorer (GPR, Stage I)

3.2.1.1 AutoEncoder (AE)

The AutoEncoder (AE) framework plays a crucial role in our IGCP architecture, functioning as a bridge that processes and refines degraded images. This AE network can be broadly categorized into two main parts: the Encoder and the Decoder. Both the encoder and decoder comprise a series of cascading residual blocks introduced in [111], which are visualized in Figure 3.3 (a).

- **Encoder:**

The encoder segment of the AE is responsible for mapping LQ images to a condensed latent space. Specifically, for a given input LQ image l , the encoder function Q processes it to produce feature maps $\mathbf{f}$ and a latent vector $\mathbf{z}$ as expressed in equation 3.12.

$$(\mathbf{f}, \mathbf{z}) = Q(l). \quad (3.12)$$

The feature maps ($\mathbf{f}$) are sent to the decoder through skip connections. The latent vector is sent to both the decoder and the Identity-Embedded StyleGAN Network (IES). The Encoder's primary mission is to find a mapping function between the LQ image and a latent vector that maintains the information needed to reconstruct the corresponding ground-truth image.

- **Decoder:**

Following the encoding process, the latent vector undergoes decoding. The decoder network is expressed in equation 3.13.

$$(\hat{\mathbf{f}}, \hat{\mathbf{h}}) = P(\mathbf{z}, \mathbf{f}) \quad (3.13)$$

where P denotes the decoder function. This function ingests the latent vector $\mathbf{z}$ and feature maps $\mathbf{f}$ produced by the encoder. The outputs are $\hat{\mathbf{f}}$, which are the skipped feature maps forwarded to the IES network, and $\hat{\mathbf{h}}$, a coarsely restored HQ image. The decoder is designed to meet the following requirements: a) Ensuring the latent code $\mathbf{z}$ is a true representative of the LQ image by attempting to construct a coarsely restored image, and b) Extracting feature maps that are then forwarded to subsequent GAN blocks in the IES network.

In a holistic view, the Encoder network maps the corrupted image to the low-dimensional latent space, while the Decoder network maps the latent representation back to the image space.

Therefore, the Encoder network conducts the GAN inversion task by providing the latent representation of the image, while the Decoder network can remove the uncorrelated noise and artifacts providing a coarsely restored image.

3.2.1.2 Identity embedded StyleGAN (IES)

The IES network, illustrated in Figure 3.3 (b), operates in conjunction with the AE network to produce the initial restoration. The IES encompasses a mapping network M and a generator G with a sequence of generative blocks.

The latent vector $\mathbf{z}$ is first projected onto an intermediate identity-infused style vector $\mathbf{s}$ through a fully connected mapping network M , mathematically presented as:

$$\mathbf{s} = M(\mathbf{z}). \quad (3.14)$$

This identity-infused style vector $\mathbf{s}$ resides within a less-entangled space [73, 76] and is trained to retain high mutual information in alignment with the identity derived from the ground-truth image, enhancing the identity preservation capability of the framework (see section 3.1.1). Subsequently, the style vector $\mathbf{s}$ and AE’s multi-resolution feature maps $\hat{\mathbf{f}}$ are fed into all generative blocks of the Generator module, formalized by:

$$\tilde{\mathbf{h}} = G(\mathbf{s}, \hat{\mathbf{f}}). \quad (3.15)$$

Each generative block (shown in Figure 3.3.(c)) consists of two consecutive style-modulated convolutions added to two different noise inputs. Notably, to embed the AE to the GAN prior network in IGCP, a constant pair of noise inputs is scaled and reused across all generative blocks.

The input of each generative block is scaled up by a factor of two until the desired output resolution is obtained. To help the generative blocks generate the desired output image, the upsampled inputs are concatenated with the corresponding skipped feature maps, where a 1×1 convolution is adopted as a bottleneck layer to keep the dimensions consistent with the original StyleGAN architecture [76].

It is worth mentioning that due to the use of a constant noise the image restoration process is a one-to-one mapping, mitigating the uncertainty problem of one-to-many mapping in some previous methods. Noise inputs are particularly important to generate fine details such as skin texture, hair strands, or subtle variations in lighting and shading. These details are often too small or too random to be controlled directly by the latent code but are essential for the realism of the images.

3.2.2 Codebook Prior Restorer (CPR, Stage II)

To further improve the image quality after the first stage, we apply the CPR network to the output of Stage (I). The CPR network is composed of a pre-trained vector quantized autoencoder [83], denoted by AE_q , that removes any remaining artifacts and distortions, and an FTM block that enhances the fidelity of the image.

3.2.2.1 Vector Quantized AutoEncoder (VQAE)

The pre-trained VQAE network [83], depicted in Figure 3.3.(d), consists of three key components: an encoder (E_q), a decoder (D_q), and a discrete codebook ($\mathcal{Z} = \{\mathbf{z}_k\}_{k=1}^K$) with $K = 1024$ entries. The encoder transforms the image $\tilde{\mathbf{h}}$ into a code denoted as $\mathbf{z}_c$ in a continuous space shown in equation 3.16:

$$\mathbf{z}_c = E_q(\tilde{\mathbf{h}}). \quad (3.16)$$

Subsequent to this encoding, the continuous codes are subjected to a nearest-neighbor operation $\mathcal{Q}(\cdot)$ to derive their corresponding quantized codes, denoted as $\mathbf{z}_q$:

$$\mathbf{z}_q = \mathcal{Q}(\mathbf{z}_c) = \underset{\mathbf{z}_k \in \mathcal{Z}}{\operatorname{argmin}} \|\mathbf{z}_c - \mathbf{z}_k\|_2^2. \quad (3.17)$$

Finally, the quantized codes are mapped back to the image space by the decoder to form the final restored image ($\mathbf{h}$):

$$\mathbf{h} = D_q(\mathbf{z}_q). \quad (3.18)$$

The VQAE network is pre-trained, following the approach in [83], utilizing both image-level and codebook-level loss terms. The image-level loss contains a combination of L1 loss, perceptual loss [112], and adversarial loss in [83], while the codebook-level loss is simply the ℓ_2 loss between the quantized code and the continuous code [83].

3.2.2.2 FTM Block

Using the VQAE network enhances image quality, but the network’s discrete latent space may compromise image fidelity and hinder identity preservation. This is primarily because mapping to a discrete space via a nearest-neighbor operation can lead to information loss. To address the pitfalls of nearest-neighbor quantization, we introduce a feature transformation module (FTM). This module aims to integrate encoder features with decoder features across different scales seamlessly (see Figure 3.3.(e)).

The FTM consists of a single residual block [111] tailored for each feature scale. It takes advantage of multi-resolution feature maps from both the encoder and decoder to derive parameters for spatial feature transformation, namely the weight $\mathcal{W}$ and shift $\mathcal{S}$:

$$\mathcal{W}, \mathcal{S} = \text{FTM}(\mathbf{f}_e^i, \mathbf{f}_d^i) \quad (3.19)$$

Here, $\mathbf{f}_e^i$ and $\mathbf{f}_d^i$ represent the encoder and decoder feature maps at the i^{th} resolution, respectively, with $\mathcal{W}, \mathcal{S}, \mathbf{f}_e^i, \mathbf{f}_d^i \in \mathbb{R}^{h \times w \times c}$, where h , w , and c denote the height, width, and number of channels.

Subsequently, the decoder feature maps undergo an update based on $\mathcal{S}$ and $\mathcal{W}$:

$$\hat{\mathbf{f}}_d^i \leftarrow \mathbf{f}_d^i \odot \mathcal{W} + \mathcal{S} \quad (3.20)$$

where $\hat{\mathbf{f}}_d^i$ denotes the the updated decoder feature at the i^{th} resolution. Such a strategy has proven effective in preserving fidelity from the encoder to the decoder without inducing distortions.

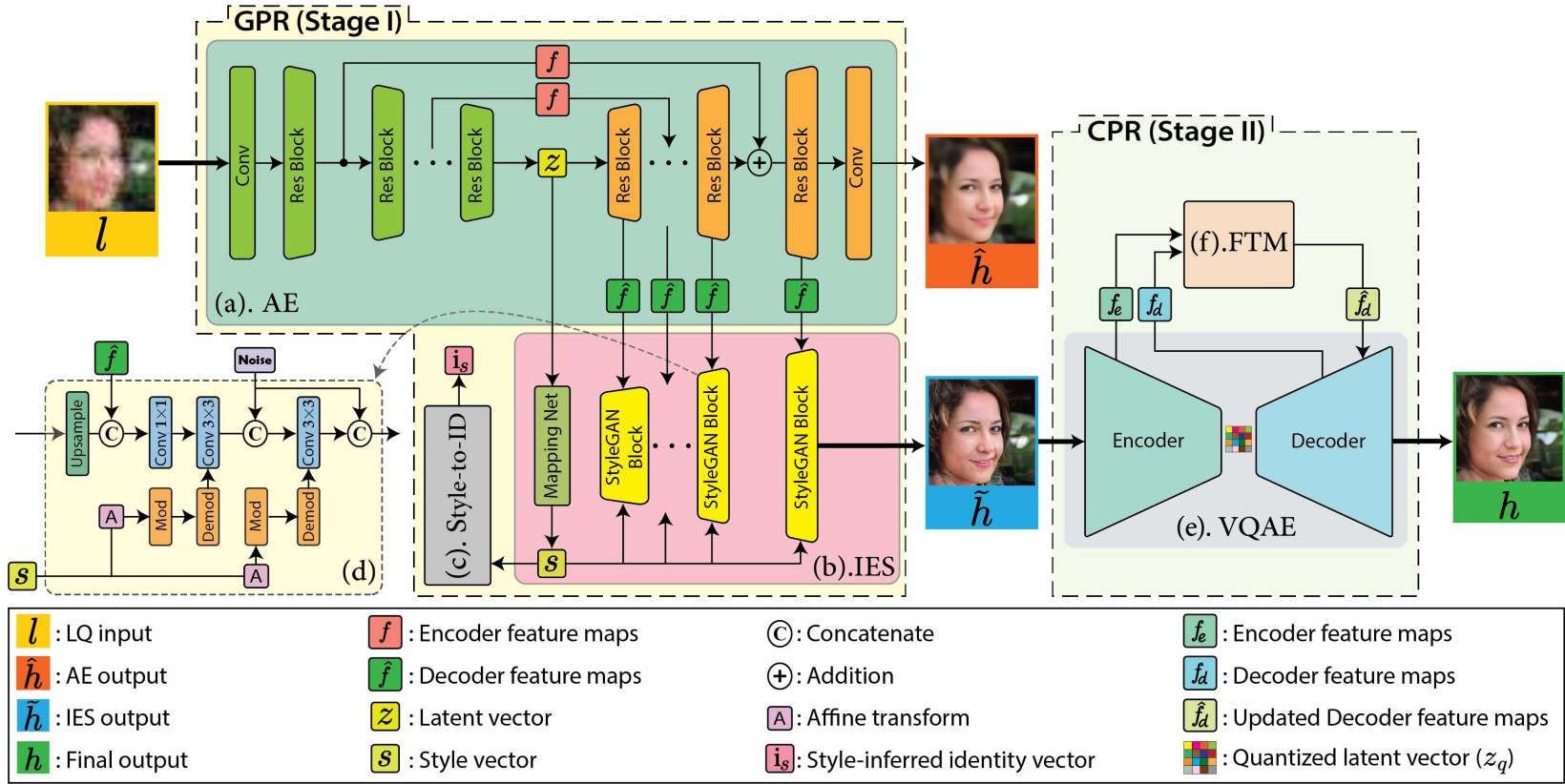


Figure 3.3: IGCP detailed architecture: (a) AE network, (b) AES Network, (c) Style-to-ID network to infer the identity vector corresponding to a style vector. (d) GAN block architecture. MOD and DEMOD are part of the weight demodulation process — an advancement from StyleGAN v2 [76] to mitigate droplet artifacts produced in the previous version. First, the modulation and convolution operations are combined by directly scaling the convolution weights (MOD). Next, output feature maps are scaled and undergo a convolution operation (DEMOD). (e) Pre-trained VQAE Network, comprising an encoder and a decoder, with weights sourced from [18]. (f) FTM Network. This lightweight module specifically designed to efficiently transfer features from the encoder to the decoder within the VQAE network.

3.3 Training Scheme

We first train the GPR network independently using a set of synthesized LQ-HQ face image pairs from the FFHQ dataset [113]. Then, having the GPR network frozen, we embed the CPR network and perform a slight fine-tuning. Figure 3.4 illustrates the training losses used in training the GPR and CPR networks:

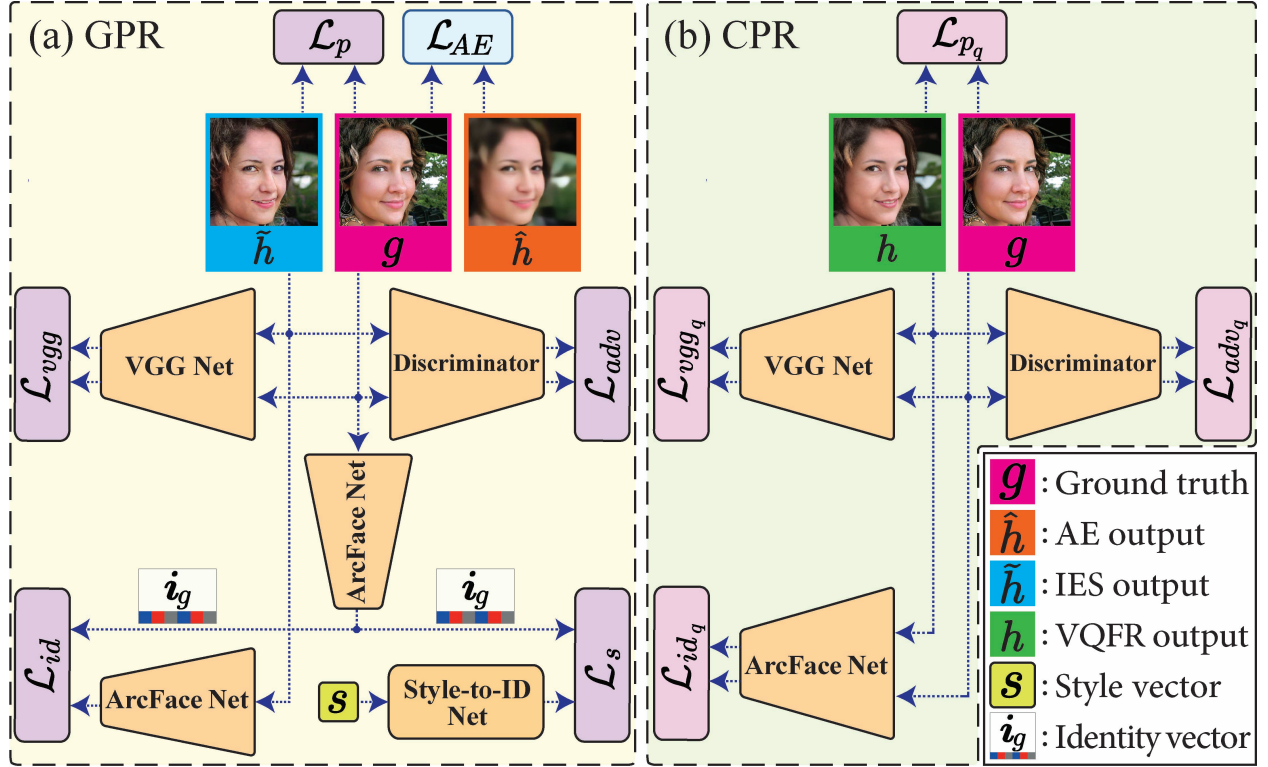


Figure 3.4: IGCP Training scheme: (a) Training loss used for the GPR network. (b) Training loss of the CPR Network.

3.3.1 GAN Prior Restorer (GPR, Stage I)

3.3.1.1 AutoEncoder (AE)

The AE network is composed of an encoder and a decoder that are connected by skip connections, forming a U-shaped structure as shown in Figure 3.3(a). The encoder aims to find an abstract representation of the LQ image in the latent space, which is the latent vector z . The decoder aims to generate feature maps that closely match the original ground-truth image and preserve the facial structure. To achieve this, we use a pixel-wise ℓ_1 loss ($\mathcal{L}_{AE}$) between the coarsely restored HQ image $\hat{h}$ produced by the decoder and the corresponding

ground-truth image. Denoting the LQ image as " l " and the ground-truth image as $\mathbf{g}$, the AE loss function can be expressed as:

$$\mathcal{L}_{AE} = \|\text{AE}(\mathbf{l}) - \mathbf{g}\|_1 \quad (3.21)$$

where $\|\cdot\|_1$ denotes the pixel-wise ℓ_1 norm applied to the AE's output, and the ground-truth image ($\mathbf{g}$).

3.3.1.2 Identity embedded StyleGAN (IES)

The IES network, depicted in Figure 3.3 (b), is a StyleGAN network⁹ that is modified to operate in conjunction with the AE network to generate the initial restoration. As for the IES network, the following four loss functions are adopted:

- a) We leverage pixel-wise ℓ_1 loss (referred to as $\mathcal{L}_p$) to fit the intermediate restored HR image $\tilde{\mathbf{h}}$ with the ground-truth image $\mathbf{g}$ and increase the fidelity. The ℓ_1 loss focuses on maintaining pixel-level accuracy:

$$\mathcal{L}_p = \|\tilde{\mathbf{h}} - \mathbf{g}\|_1. \quad (3.22)$$

- b) We leverage a VGG-based perceptual loss [112] (denoted as $\mathcal{L}_{vgg}$) to maintain perceptual quality. The VGG loss emphasizes perceptual similarity by considering high-level feature representations extracted by the VGG network [101]. The formula for this loss is given by:

$$\mathcal{L}_{vgg} = \sum_{j=1}^N \frac{1}{C_j H_j W_j} \left\| \phi_j(\tilde{\mathbf{h}}) - \phi_j(\mathbf{g}) \right\|_2^2 \quad (3.23)$$

where the function $\Phi_j(\cdot)$ represents the feature map obtained from the j^{th} layer of a pre-trained VGG network for a given input. The dimensions of these feature maps are denoted by C_j , H_j , and W_j (representing the number of channels, height, and width, respectively). The parameter N corresponds to the number of VGG layers used in the loss computation. The term $\|\cdot\|_2^2$ signifies the squared Euclidean (ℓ_2) norm. This loss function aims to minimize the difference in high-level features between the target and generated images, encouraging perceptual similarity in terms of content and texture.

⁹ The initial weights of the StyleGAN network were from a pre-trained model, providing "priors". The network is fine-tuned during the training of our model.

- c) We also incorporate the adversarial loss (referred to as $\mathcal{L}_{adv}$) within the adversarial framework to ensure residing on the natural image manifold:

$$\mathcal{L}_{adv} = \log(D(\mathbf{g})) + \log(1 - D(\tilde{\mathbf{h}})) \quad (3.24)$$

where $D(\cdot)$ denotes the discriminator network.

- d) Similar to GFPGAN method [16], we employ an *identity-in-image* loss to preserve facial identities. This loss measures the cosine distance between the ground-truth identity ($\mathbf{i}_g$) and the inferred identity from the output image ($\mathbf{i}_h$) as follows:

$$\mathcal{L}_{id} = 1 - \cos(\mathbf{i}_g, \mathbf{i}_h) \quad (3.25)$$

where the identity vectors are obtained using the ArcFace network [33], commonly employed for face recognition tasks.

- e) To enhance the preservation of identity information during the image generation process, we propose the *identity-in-style* loss denoted as $\mathcal{L}_s$, which encourages a high mutual information between the identity of the images and their corresponding styles ensuring the latent codes belonging to the same identities reside close to each other in the style space. To achieve this objective, We utilize a fully connected network with a single hidden layer, referred to as the Style-to-ID network, to approximate the mean vector ($f(\mathbf{s})$). This nomenclature underscores the network's purpose of mapping style to identity.

$$\mathcal{L}_s = \|\mathbf{i}_g - f(\mathbf{s})\|^2. \quad (3.26)$$

The total loss of the IES network is the combination of all the mentioned loss functions:

$$\mathcal{L}_{IES} = \mathcal{L}_p + \mathcal{L}_{vgg} + \mathcal{L}_{adv} + \mathcal{L}_{id} + \mathcal{L}_s. \quad (3.27)$$

It should be noted that the AE network undergoes simultaneous training in tandem with the IES network. However, gradients do not flow from the AE to the IES, thus not directly impacting StyleGAN's fine-tuning. This decoupling of the gradients allows the network to focus on learning distinct aspects separately, leading to better results and more efficient training. By preserving identity in the style vectors, we ensure that the generated

images possess identities that are more consistent with the ground-truth images. This approach strengthens the network’s ability to produce faithful representations and maintain the distinctive characteristics of the original faces.

The overall loss function for the GPR network, denoted as $\mathcal{L}_{GPR}$, is formed by aggregating $\mathcal{L}_{AE}$ (3.21) and $\mathcal{L}_{IES}$ (3.27):

$$\mathcal{L}_{GPR} = \mathcal{L}_{AE} + \mathcal{L}_{IES}. \quad (3.28)$$

3.3.2 Codebook Prior Restorer (CPR, Stage II)

In the second stage, we refine the image quality and eliminate any remaining artifacts and distortions from the first stage by feeding the image restored in Stage (I) to the CPR network. The CPR network comprises a VQAE network for distortion removal and an FTM block for fidelity enhancement.

We employ a VQAE network, which has been pre-trained based on the methodology described in [83]¹⁰. This network utilizes both image-level and codebook-level loss functions. Specifically, the codebook-level loss is determined by the ℓ_2 loss between the quantized code and its continuous counterpart, as detailed in [83]. However, given that the Encoder network (E_q) remains fixed in our approach, we opt to disregard the codebook-level loss, focusing solely on the image-level loss which amalgamates the following loss functions:

- a) $\mathcal{L}_{pq}$, a pixel-wise ℓ_1 loss between the final restored image $\mathbf{h}$ and the ground-truth image $\mathbf{g}$ as shown in Equation 3.29:

$$\mathcal{L}_{pq} = \|\mathbf{h} - \mathbf{g}\|_1. \quad (3.29)$$

- b) $\mathcal{L}_{vggq}$, a perceptual loss [112, 114] defined as the perceptual dissimilarity between the final restored image $\mathbf{h}$ and the ground-truth image $\mathbf{g}$ by considering high-level feature representations extracted by the VGG network [101]:

$$\mathcal{L}_{vggq} = \sum_{j=1}^N \frac{1}{C_j H_j W_j} \left\| \phi_j(\tilde{\mathbf{h}}) - \phi_j(\mathbf{g}) \right\|_2^2 \quad (3.30)$$

¹⁰ The initial weights for this VQAE network are sourced from [18].

where the function $\Phi_j(\cdot)$ represents the feature map obtained from the j^{th} layer of a pre-trained VGG network for a given input. The dimensions of these feature maps are denoted by C_j , H_j , and W_j (representing the number of channels, height, and width, respectively). The parameter N corresponds to the number of VGG layers used in the loss computation. The term $\|\cdot\|_2^2$ signifies the squared Euclidean (ℓ_2) norm. This loss function aims to minimize the difference in high-level features between the target and generated images, encouraging perceptual similarity in terms of content and texture.

c) $\mathcal{L}_{adv_q}$, the adversarial loss in [83] as shown in Equation 3.31.

$$\mathcal{L}_{adv_q} = \log(D_q(\mathbf{g})) + \log(1 - D_q(\mathbf{h})) \quad (3.31)$$

where $D_q(\cdot)$ denotes the discriminator network in the CPR module.

In addition to the loss functions mentioned above, we also use an identity-in-image loss in Equation 3.32 that measures the cosine similarity between the identity vectors of the final restored image $\mathbf{i}_h$ and the ground truth image $\mathbf{i}_g$, obtained using the ArcFace network [33]. This inclusion aims to encourage the FTM module to effectively convey identity-rich features.

$$\mathcal{L}_{id_q} = 1 - \cos(\mathbf{i}_h, \mathbf{i}_g). \quad (3.32)$$

The overall loss function for the CPR network, denoted as $\mathcal{L}_{CPR}$, is formed by aggregating the above mentioned loss functions as shown in Equation 3.33:

$$\mathcal{L}_{CPR} = \mathcal{L}_{pq} + \mathcal{L}_{vgg_q} + \mathcal{L}_{adv_q} + \mathcal{L}_{id_q}. \quad (3.33)$$

The $\mathcal{L}_{CPR}$ loss aids in optimizing parameters within the FTM module. We also slightly fine-tune the VQAE network along with the FTM training.

In summary, our proposed two-stage approach combines the strengths of a U-shaped autoencoder and StyleGAN, along with the benefits of quantized latent spaces from VQGAN, to progressively restore degraded face images. By addressing both coarse and fine-grained details and preserving facial identities throughout the restoration process, we aim to provide a robust and effective blind face restoration method capable of producing highly realistic and high-fidelity outputs.

Chapter 4

Experiments

4.1 Datasets

4.1.1 Training Dataset.

We use the FFHQ dataset [73], comprising 70,000 HQ face images of resolution 1024×1024 . The ground-truth images ($\mathbf{g}$) are formed by resizing FFHQ images to 512×512 . For training pairs, we derive LQ ($\mathbf{l}$) images from their ground-truth counterparts using the following degradation model [16, 17, 18, 19]:

$$\mathbf{l} = \left\{ [\mathbf{g} \otimes k_\sigma \downarrow_r + n_\delta]_{\text{JPEG}q} \right\} \uparrow_r \quad (4.1)$$

where $\mathbf{g}$ is first filtered with a Gaussian kernel k_σ , downsampled by factor r , added with Gaussian noise n_δ , compressed using JPEG with quality q , and finally resized to 512×512 . The parameters σ , r , δ , and q are uniformly sampled from $[0.2:10]$, $[1, 16]$, $[0, 15]$, and $[60, 100]$ intervals, respectively.

Additionally, for data augmentation purposes, each image in the training set is independently subjected to random transformations in each epoch. Specifically, with a 20% probability, an image is mirrored, and with a 10% probability, it is converted to black and white. Figure 4.1 presents example pairs of HQ images and their corresponding LQ counterparts from our training set.

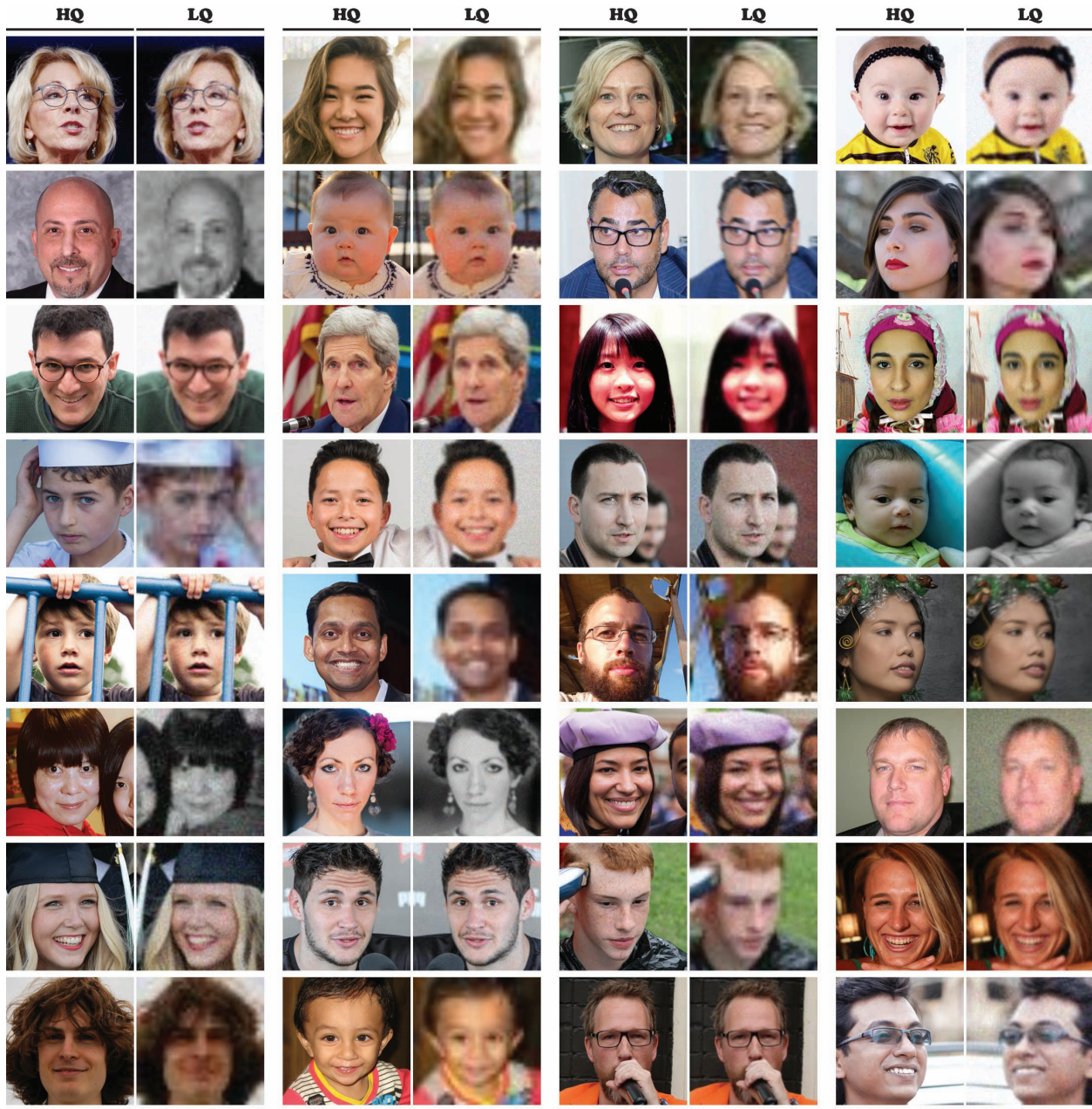


Figure 4.1: Example pairs of HQ ground-truth images from the FFHQ dataset [73], alongside their LQ counterparts derived using Equation 4.1.

4.1.2 Test Datasets.

4.1.2.1 Synthetic datasets

We evaluated our trained model on two synthetic datasets, where the LQ input images are derived from their HQ ground-truth counterparts, forming pairs of LQ-HQ images.

- (i) **CelebA-IntenseTest**, comprising 3,000 LQ-HQ image pairs derived from CelebA-HQ [86]. The LQ images underwent high degradation using the Equation 4.1, where the parameters σ , r , δ , and q are uniformly sampled from [5:20], [2, 16], [0, 20], and [50, 80] intervals, respectively. Figure 4.2 presents example pairs of HQ images and their corresponding LQ counterparts from the CelebA-IntenseTest dataset.

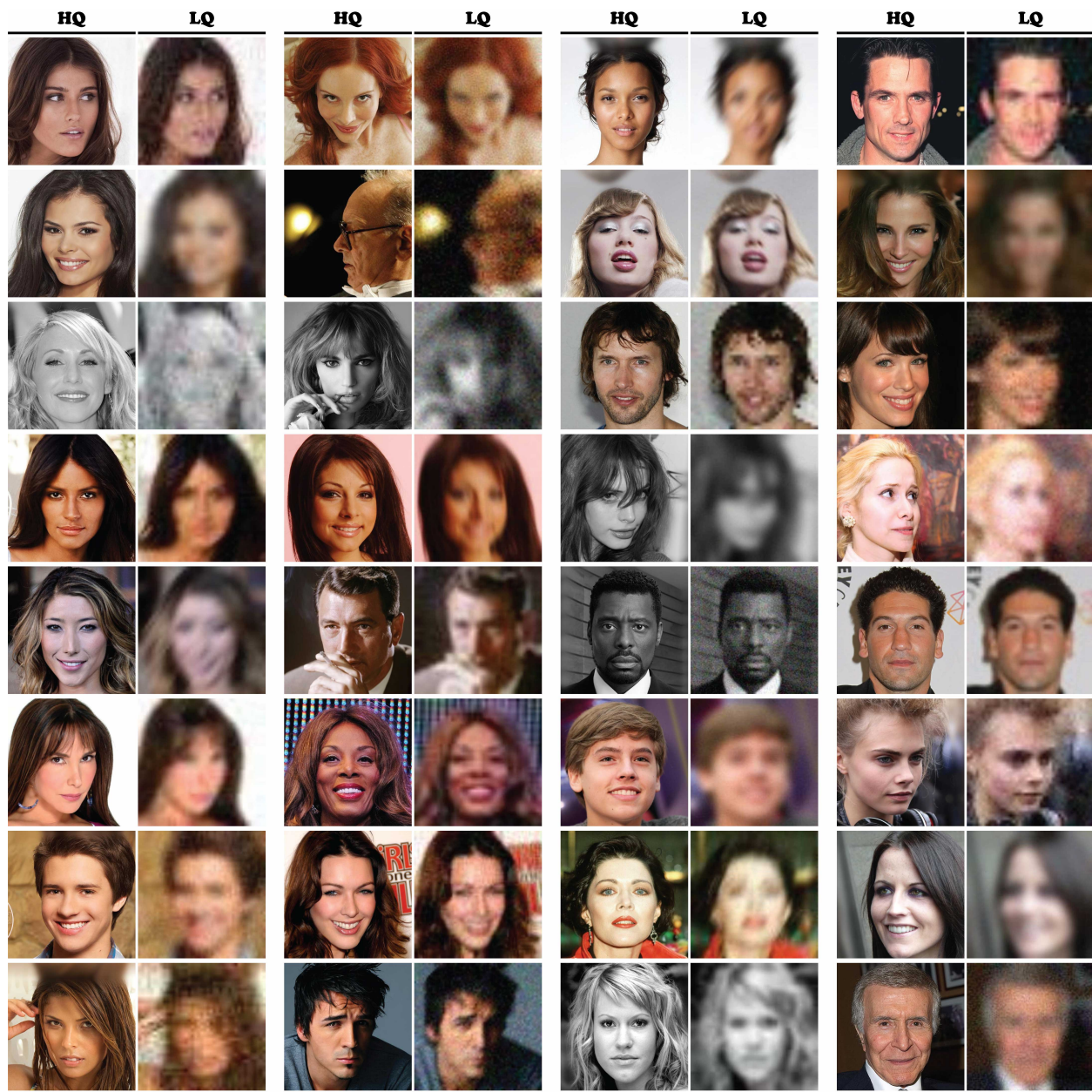


Figure 4.2: CelebA-IntenseTest dataset: example pairs of HQ ground-truth images and their synthesized LQ counterparts.

- (ii) **CelebA-Test** proposed by [16]. This dataset consists of 3,000 LQ-HQ image pairs derived from CelebA-HQ [86]. The LQ images are generated by applying a moderate degradation process ¹¹ to the HQ images, which makes them less challenging than the ones in CelebA-IntenseTest dataset. Figure 4.3 presents example pairs of HQ images and their corresponding LQ counterparts from the CelebA-Test dataset.

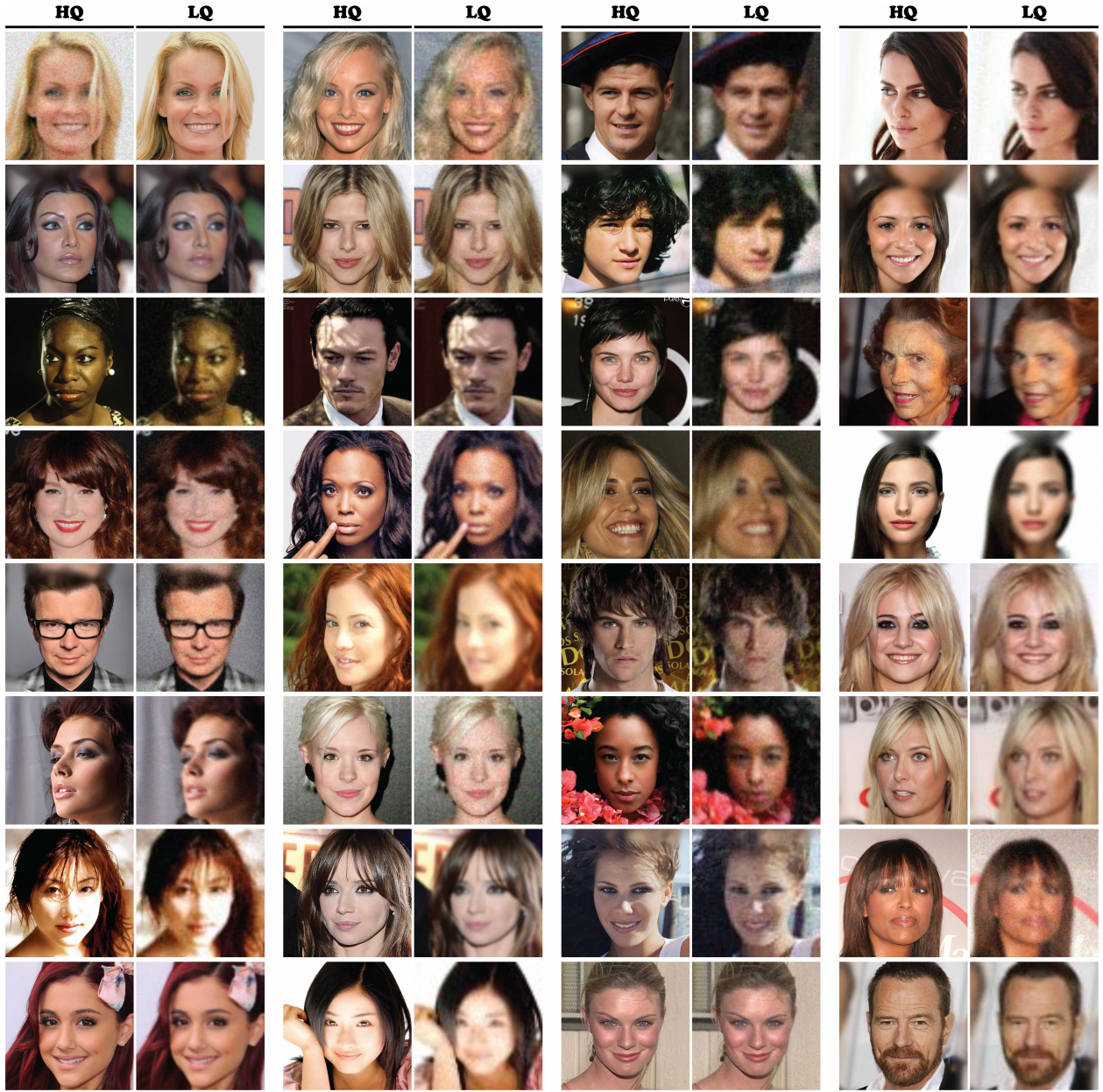


Figure 4.3: CelebA-Test dataset: example pairs of HQ ground-truth images and their synthesized LQ counterparts.

¹¹ the parameters σ , r , δ , and q in Equation 4.1 are randomly selected from $[0.2:10]$, $[1, 8]$, $[0, 15]$, and $[60, 100]$, respectively [16].

4.1.2.2 real-world datasets

We also assessed the model’s performance on three real-world datasets: LFW-Test, WebPhoto (proposed by [16]), and Wider-Test [18]. These datasets feature only LQ images with diverse degradations, and their ground-truth HQ counterparts are unknown. Furthermore, it should be noted that they do not overlap with our training set.

- (i) **LFW-Test** [16] (*degradation level: mild*). The LFW-Test dataset is a subset of the LFW dataset [88], which consists of LQ face images taken in the wild. It includes only the first image of each individual in the LFW dataset, resulting in a total of 1,711 images. Figure 4.4 presents some example images from the LFW-Test dataset.



Figure 4.4: Representative examples from the LFW-Test dataset.

- (ii) **WebPhoto-Test** [16] (*degradation level: moderate*). The WebPhoto-Test dataset comprises 407 LQ faces collected from the Internet. These faces were extracted from 188 photos that exhibit diverse and complicated degradation. Figure 4.5 displays some images from the WebPhoto-Test dataset.



Figure 4.5: Representative examples from the WebPhoto-Test dataset.

- (iii) **WIDER-Test** [18] (*degradation level: Heavy*). The WIDER-Test dataset includes 970 severely degraded face images, derived from the WIDER Face dataset [115]. These images serve as a challenging benchmark, aiming to evaluate the generalizability and robustness of blind face restoration methods. Figure 4.6 showcases some images from the WIDER-Test dataset.

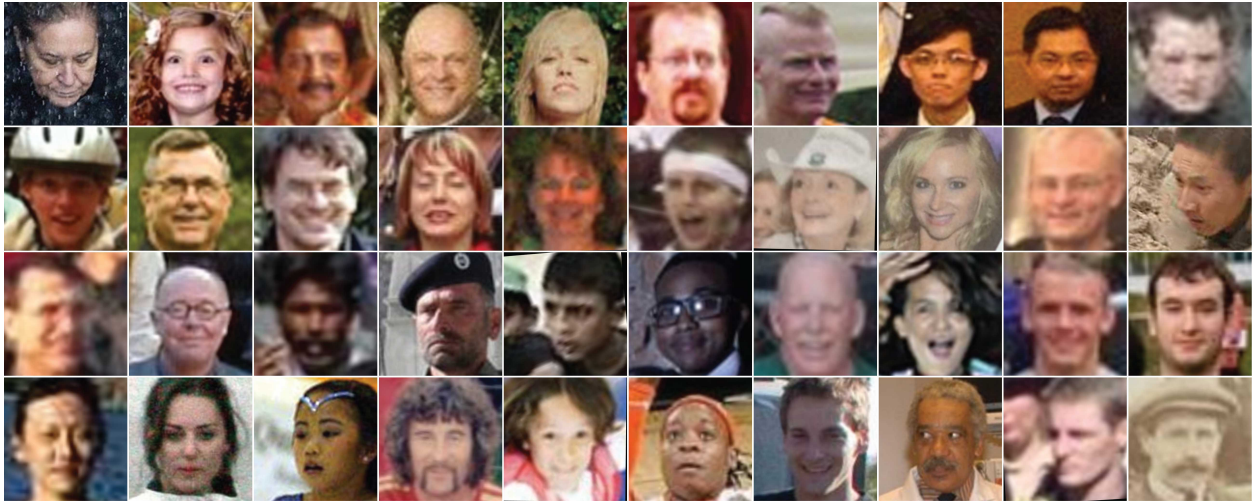


Figure 4.6: Representative examples from the WIDER-Test dataset.

4.2 Metrics

To evaluate the proposed method on CelebA-HQ with available ground-truth, we utilize reference-based measures, **PSNR**, Structural Similarity Index Measure (**SSIM**) [30], and (**LPIPS**) [32]. Additionally, we assess identity preservation by employing the cosine similarity of features extracted from the ArcFace network [33], denoted as **IDS**. Furthermore, we evaluate the perceptual realism of the restored images across both the CelebA and real-world datasets, by Frechet Inception Distance (**FID**) [31] metric.

4.3 Implementation Details

The GPR module’s AE network comprises a convolution layer at both ends, complemented by 6 cascaded residual blocks [111] in the encoder and decoder. Each convolution is followed by batch normalization and Leaky ReLU activation (negative slope of 0.2). The style-to-ID network utilizes two fully connected layers with a ReLU activation function between them. The architectures of the StyleGAN and VQFR networks follow the specifications in [76] and [83], respectively. The FTM block includes a 1×1 convolution for merging features from the VQFR’s encoder and decoder, followed by a residual block.

We implemented our method using the PyTorch framework and distributed the training across 8 GeForce GTX 1080 Ti GPUs. Our training process is divided into two distinct stages: GPR (stage I) and CPR (stage II).

- (I) **Stage I:** In this stage, we fine-tuned the IES network concurrently with the training of the AE network. Following the methodology outlined in [76], we employed the Adam optimizer [116] with a learning rate set to 10^{-3} . The hyperparameters β_1 and β_2 were set to 0.9 and 0.999 respectively. We utilized a batch size of 16 and trained the network for a total of 400k iterations.
- (II) **Stage II:** In alignment with the practices of [83], we once again employed the Adam optimizer [116], but this time with a reduced learning rate of 10^{-5} . The hyperparameters β_1 and β_2 remained consistent at 0.9 and 0.999. The batch size was reduced to 8 for this stage. The training proceeded for 10k iterations. During the first 5k iterations, we fixed the parameters of the VQAE network, only updating those of the FTM module. Subsequently, in the remaining iterations, we fine-tuned both the decoder of the VQAE network and the FTM module. While doing so, we scaled down the CPR loss, $\mathcal{L}_{CPR}$ (refer to Equation 3.33), by a factor of 0.08.

4.4 Comparison with SOTA Methods

In this section, we benchmark the performance of IGCP against several SOTA techniques: PULSE [9], GFPGAN [16], GPEN [17], DR2 [1], Codeformer [18], and VQFR [73]. The first three methods belong to GAN prior based techniques, DR2 is a diffusion-based method, and the last two methods fall under the category of codebook prior-based techniques.

All the aforementioned models were pretrained on the FFHQ dataset. For our evaluation, we ran them using their default settings to ensure a fair comparison.

4.4.1 Quantitative Comparison.

4.4.1.1 Comparison on Synthesized Datasets

The following tables provide a comparative evaluation of IGCP against SOTA baselines on two datasets: CelebA-IntenseTest, which features heavy degradation, and CelebA-Test, characterized by moderate degradation.

Table 4.1: Quantitative comparison on CelebA-IntenseTest (heavy degradation). Measures for LQ are shown as a point of reference.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	IDS $\uparrow$
LQ	21.85	0.768	0.603	172.01	0.37
PULSE	19.45	0.719	0.522	50.78	0.13
GFPGAN	21.00	0.725	0.452	27.01	0.45
GPEN	21.73	0.728	0.448	27.43	0.49
DR2	20.64	0.738	0.469	35.05	0.27
CodeFormer	22.03	0.730	0.422	30.65	0.49
VQFR	22.02	0.745	0.415	19.11	0.52
IGCP	22.67	0.759	0.404	16.40	0.59

As shown in Table 4.1, the IGCP model achieves remarkable performance on CelebA-IntenseTest, surpassing all other methods in terms of various metrics. This impressive result demonstrates that IGCP can handle severe image degradations effectively and robustly, while preserving the facial details and identity.

Table 4.2: Quantitative comparison on CelebA-Test (moderate degradation). Measures for LQ are shown as a point of reference.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	IDS $\uparrow$
LQ	25.31	0.821	0.487	101.86	0.79
PULSE	20.46	0.743	0.484	44.52	0.18
GFPGAN	24.67	0.807	0.343	11.72	0.81
GPEN	24.65	0.787	0.400	14.65	0.83
DR2	21.71	0.752	0.424	26.17	0.34
CodeFormer	25.12	0.802	0.344	14.99	0.77
VQFR	24.75	0.808	0.346	10.28	0.81
IGCP	25.24	0.813	0.348	11.61	0.84

As shown in Table 4.2, the IGCP model achieves remarkable performance compared to other models. It surpasses them in PSNR, SSIM, and IDS metrics, which reflect its ability to generate HQ images from corrupted inputs. It also performs well and achieves competitive results in LPIPS and FID, which evaluate the perceptual similarity and diversity of the generated images. These results demonstrate that IGCP can handle varying degrees of image degradations while preserving essential image characteristics.

4.4.1.2 Comparison on Real-World Datasets

Table 4.3 contrasts restoration methods on real-world datasets using the FID and IDS metrics. FID gauges the similarity between restored and real images, while IDS assesses identity preservation.

Table 4.3: Quantitative comparison on real-world datasets.

Dataset	LFW-Test		WebPhoto-Test		WIDER-Test	
Method	FID $\downarrow$	IDS $\uparrow$	FID $\downarrow$	IDS $\uparrow$	FID $\downarrow$	IDS $\uparrow$
LQ	136.80	1.00	168.36	1.00	197.46	1.00
PULSE	83.80	0.20	101.03	0.42	86.36	0.18
GFPGAN	61.25	0.93	91.11	0.82	43.08	0.74
GPEN	58.89	0.95	82.08	0.83	50.23	0.72
DR2	51.03	0.40	107.21	0.42	47.84	0.36
CodeFormer	59.32	0.89	85.35	0.77	43.32	0.70
VQFR	55.47	0.88	87.62	0.80	43.87	0.72
IGCP	55.93	0.95	80.54	0.84	43.11	0.74

The results in Table Table 4.3 reveal the following key points:

1. IGCP consistently posts commendable FID scores, signifying its adeptness at generating perceptually accurate restorations.
2. IGCP’s dominance in the IDS metric across all datasets highlights its superior ability to maintain the input image’s identity. This is especially critical in real-world scenarios, where the fidelity of individual faces is paramount.

4.4.2 Qualitative Comparison

4.4.2.1 Visual Comparison on Synthesized Datasets

Figure 4.7 provides a visual evaluation of restoration methods on CelebA-HQ images that were synthetically degraded using equation 4.1.

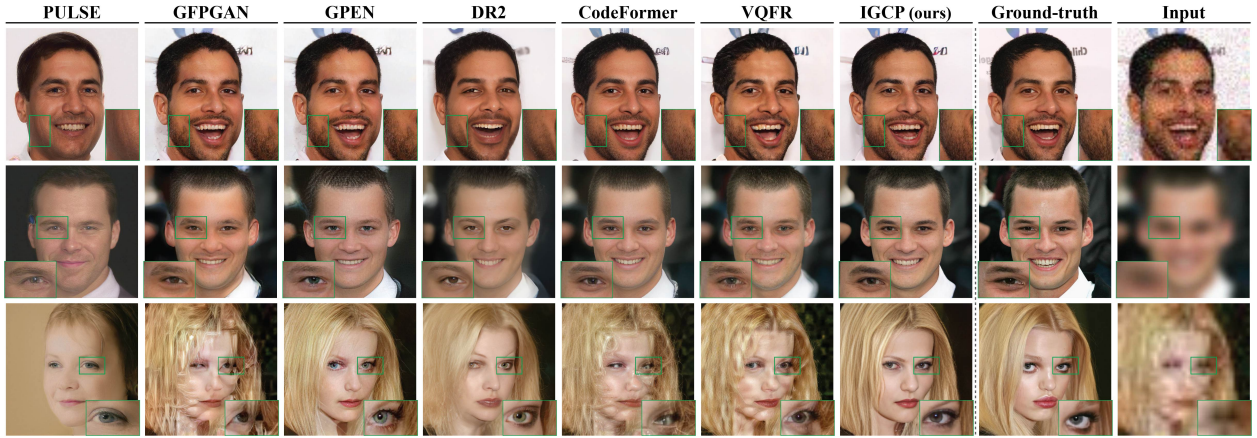


Figure 4.7: Qualitative comparison of IGCP with SOTA models on CelebA-IntenseTest dataset. GT means ground-truth.

The comparison performed in Figure 4.7 reveals the following insights:

1. *Mild and Moderate Degradation* (upper and middle rows): In cases where the facial features are not significantly distorted by degradation and the subject’s identity is still discernible, IGCP stands out by enhancing image quality and ensuring the preservation of the subject’s identity marking its advantage over other SOTA methods.
2. *Heavy Degradation* (lower row): IGCP effectively maintains image quality and retains recognizable facial features highlighting its resilience to severe degradation.

Figure 4.8 illustrates a comparative evaluation of IGCP and other SOTA models using the CelebA-IntenseTest dataset, which consists of synthesized LQ images paired with their ground-truth. Given the severe degradation in these images, most SOTA models struggle to deliver satisfactory restorations. In contrast, the IGCP model showcases significant robustness to noise. Notably, even when presented with minimal cues, IGCP effectively utilizes the available information to reconstruct the identity, producing HQ outputs exhibiting a commendable degree of similarity to the ground-truth.

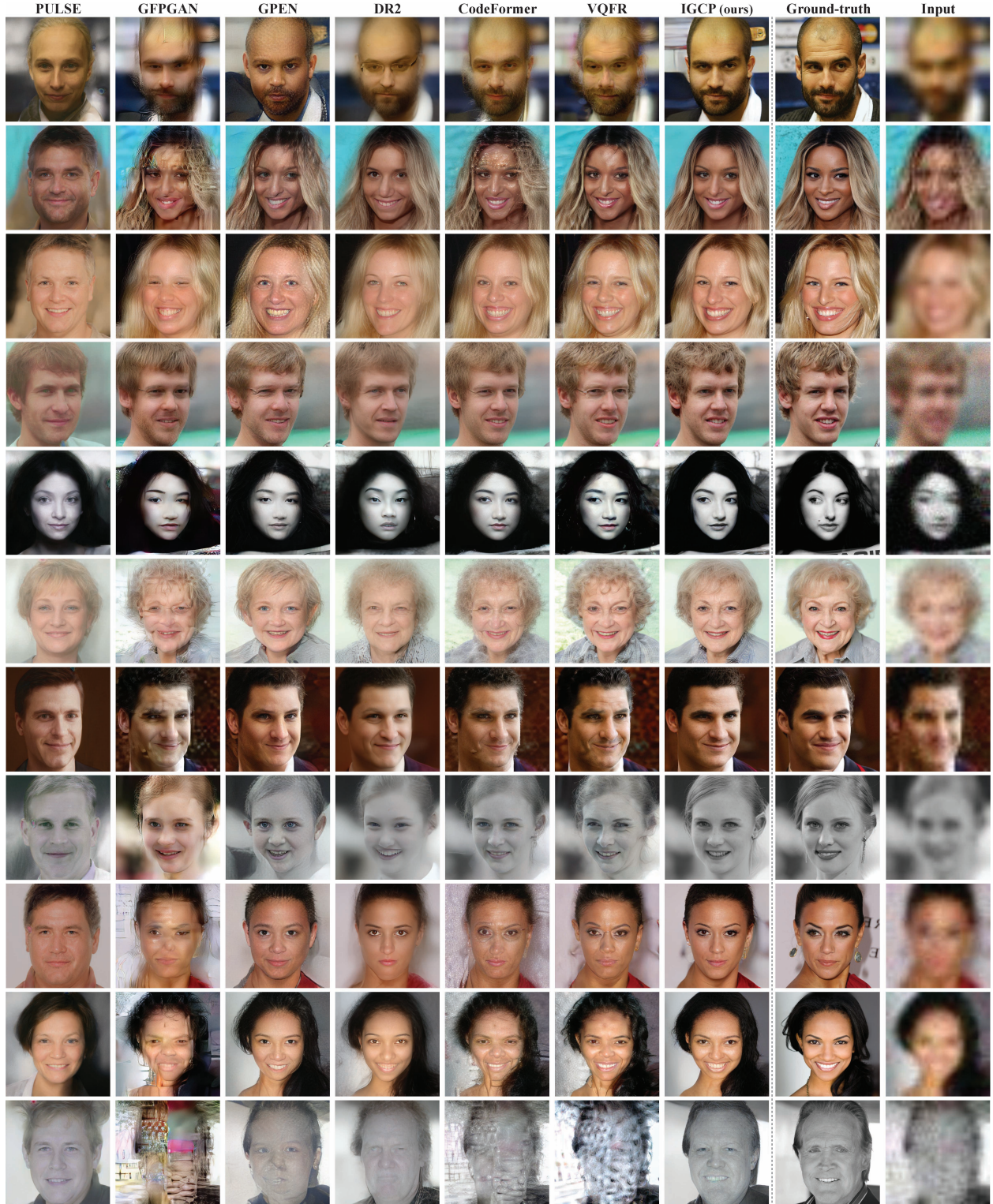


Figure 4.8: Qualitative comparison of IGCP with SOTA models on **CelebA-IntenseTest** dataset (heavy degradation).

Figure 4.9 offers a qualitative evaluation of the IGCP model’s performance in comparison to other SOTA models using the CelebA-Test dataset, characterized by mild degradation. Despite the relatively subtle distortions in the dataset, discerning the nuances and producing high-fidelity restorations can still pose a challenge. However, the IGCP model, as evident from the results, manages to restore images that bear a striking resemblance to the ground-truth. This consistency underlines the model’s prowess in handling both extreme and subtle degradations, reiterating its versatility and efficacy in diverse image restoration scenarios.

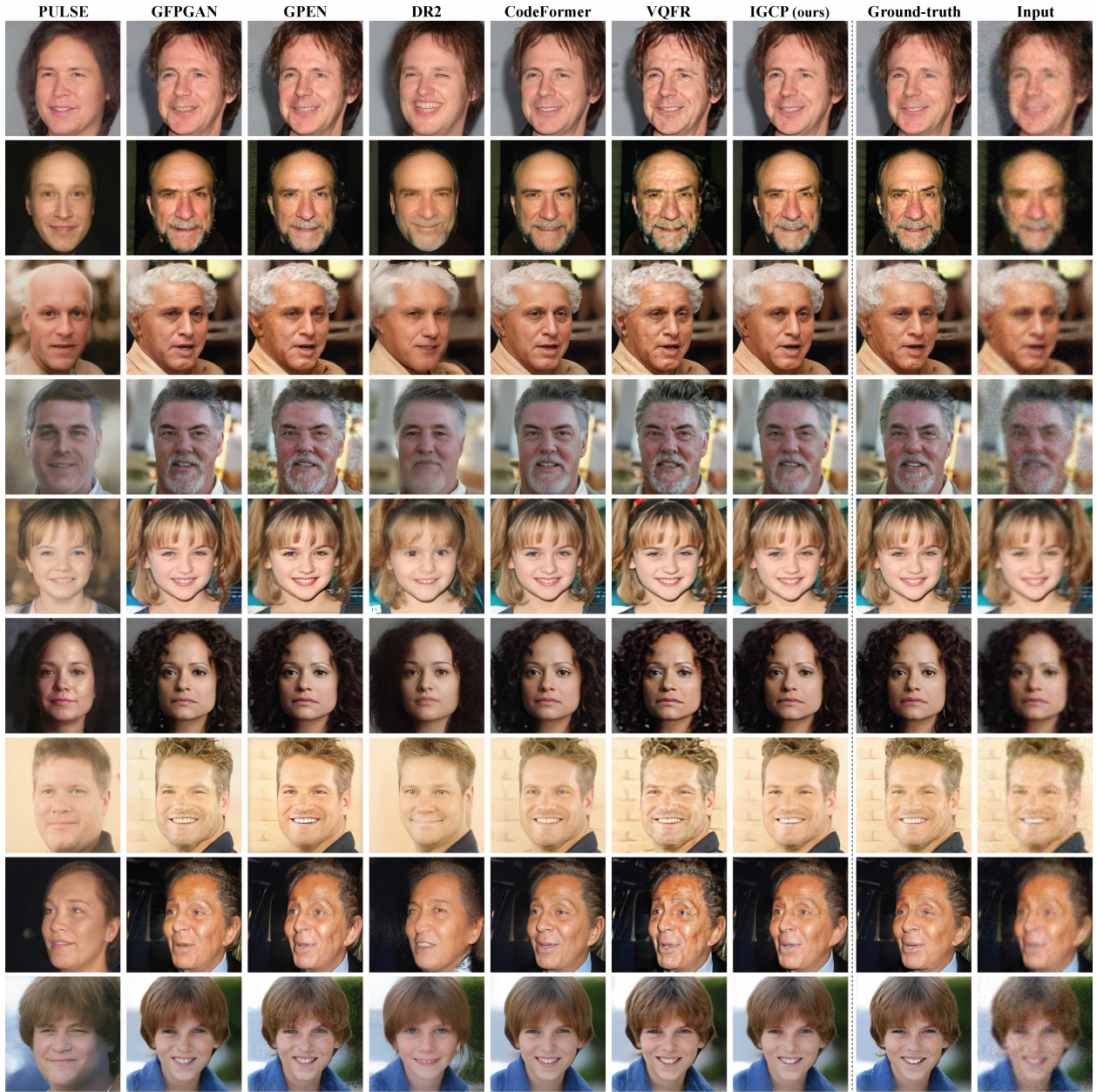


Figure 4.9: Qualitative comparison of IGCP with SOTA models on **CelebA-Test** dataset (mild degradation).

4.4.2.2 Visual Comparison on Real-World Datasets

Figure 4.10 offers a visual evaluation of BFR methods across three real-world datasets. Evidently, while most SOTA baselines grapple with reconstructing facial components in the presence of severe degradation (as illustrated in the bottom row), IGCP stands out by successfully restoring these features, underlining its advanced generalization capabilities. Further observations indicate that for images with mild to moderate degradation (top and middle rows respectively), IGCP produces sharper and more visually appealing outcomes, where facial components align more closely with the originals.

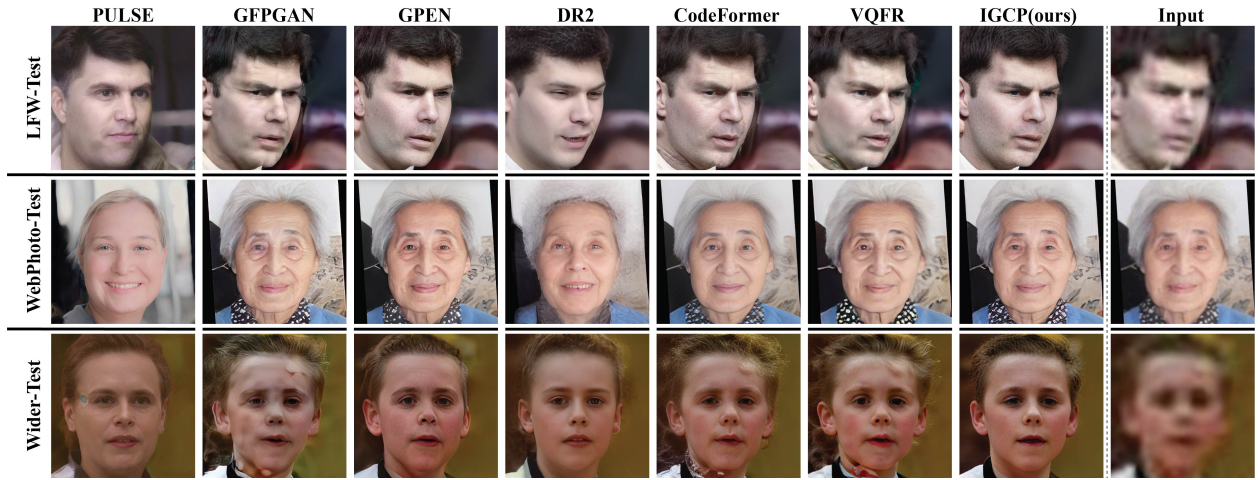


Figure 4.10: Qualitative comparison of IGCP with SOTA models on real-world datasets: LFW-Test [16], WebPhoto-Test [16], and WIDER-Test [18].

Figure 4.11 showcases a comparative analysis between IGCP and other SOTA models on the Wider-Test dataset, characterized by a high level of degradation in LQ images. A striking observation from this comparison is IGCP’s robustness against noise. Despite the high degradation level, IGCP stands out, producing HQ images that maintain a strong fidelity to the original, while other SOTA models grapple with the noise and degradation. Another significant strength of IGCP lies in its age-sensitive restoration. This is prominently displayed in the first row: when restoring an image of a child, the output from IGCP is age-appropriate, unlike other models that produce an anomalously aged appearance. On the other hand, in the second row, IGCP illustrates its capability to restore mature facial features adeptly. The restored face displays natural-looking skin wrinkles, reflecting an age-consistent skin texture.

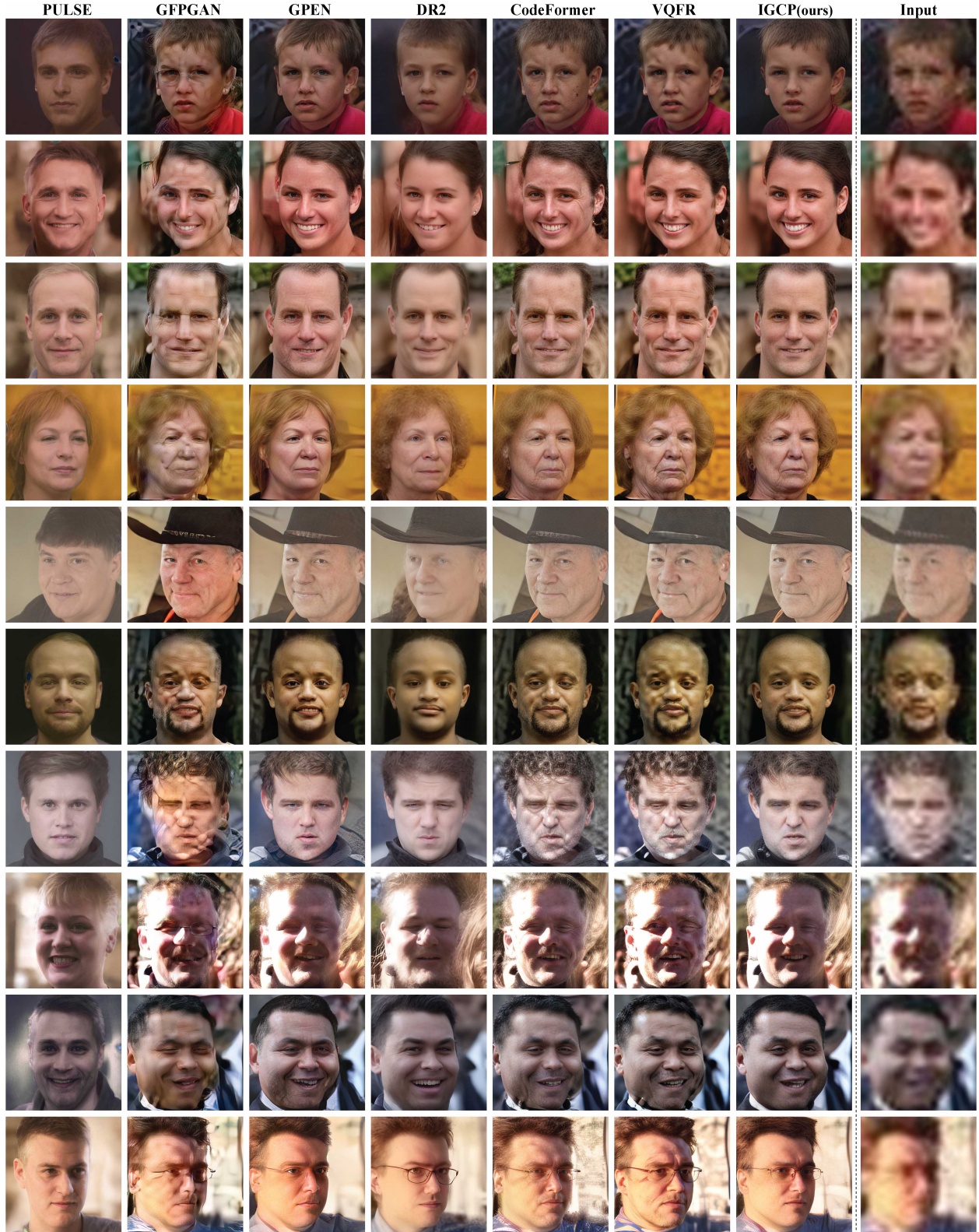


Figure 4.11: Qualitative comparison of IGCP with SOTA models on **WIDER-Test** dataset[18] (heavy degradation).

Figure 4.12 presents a qualitative comparison of the images restored using the IGCP method and other SOTA models on the WebPhoto dataset, which contains LQ images with an average level of degradation. It is evident that IGCP consistently produces HQ restorations. Importantly, this quality doesn’t come at the expense of fidelity to the original image. A prime example of this balance can be observed in the eyes of the woman in the first row: they are not only enhanced in clarity, but also remain closely aligned to the original LQ image, ensuring the preservation of key facial features. Through these results, we demonstrate IGCP’s ability to restore images without compromising the core identity and inherent attributes of the subject.

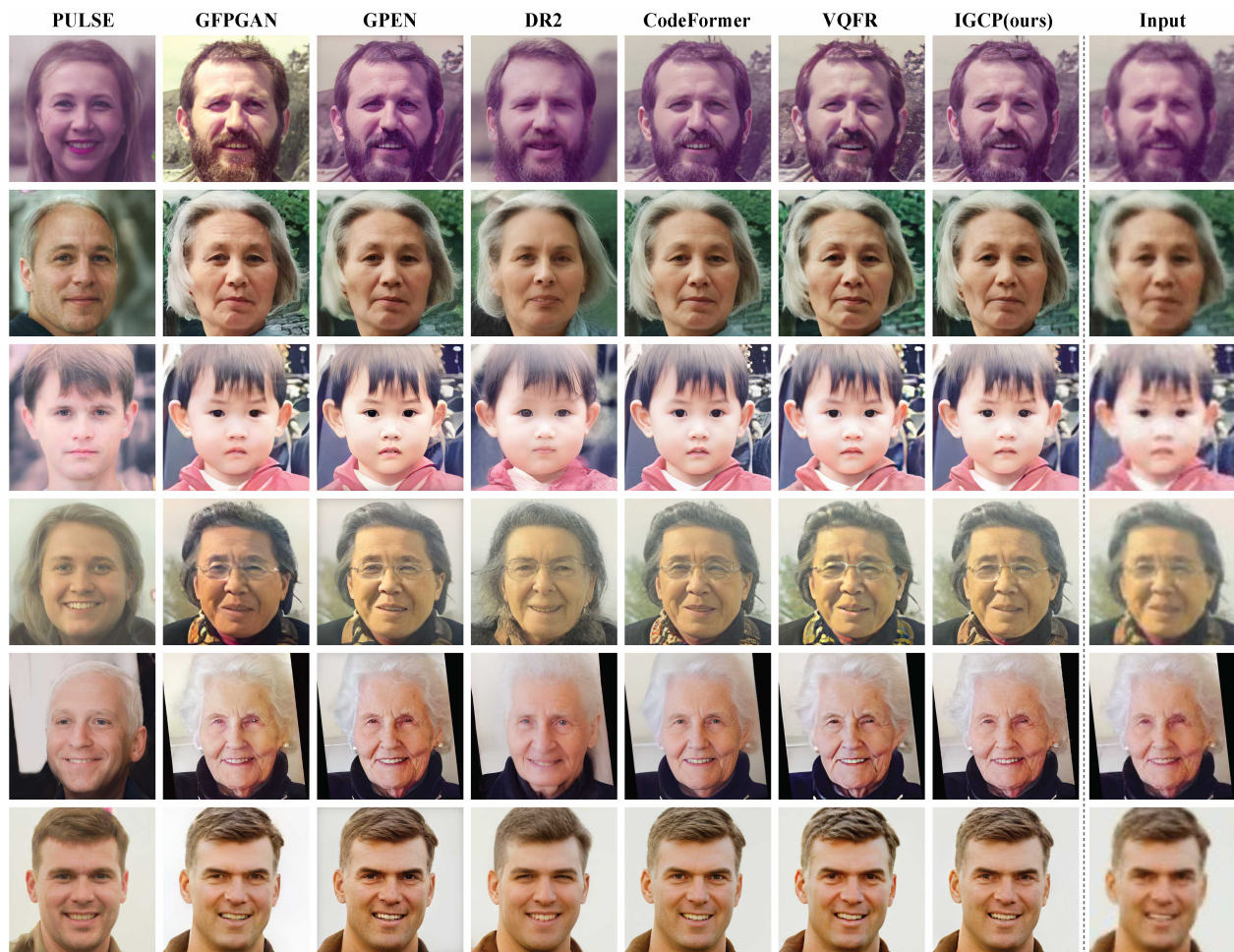


Figure 4.12: Qualitative comparison of IGCP with SOTA models on **WebPhoto-Test** dataset [16] (moderate degradation).

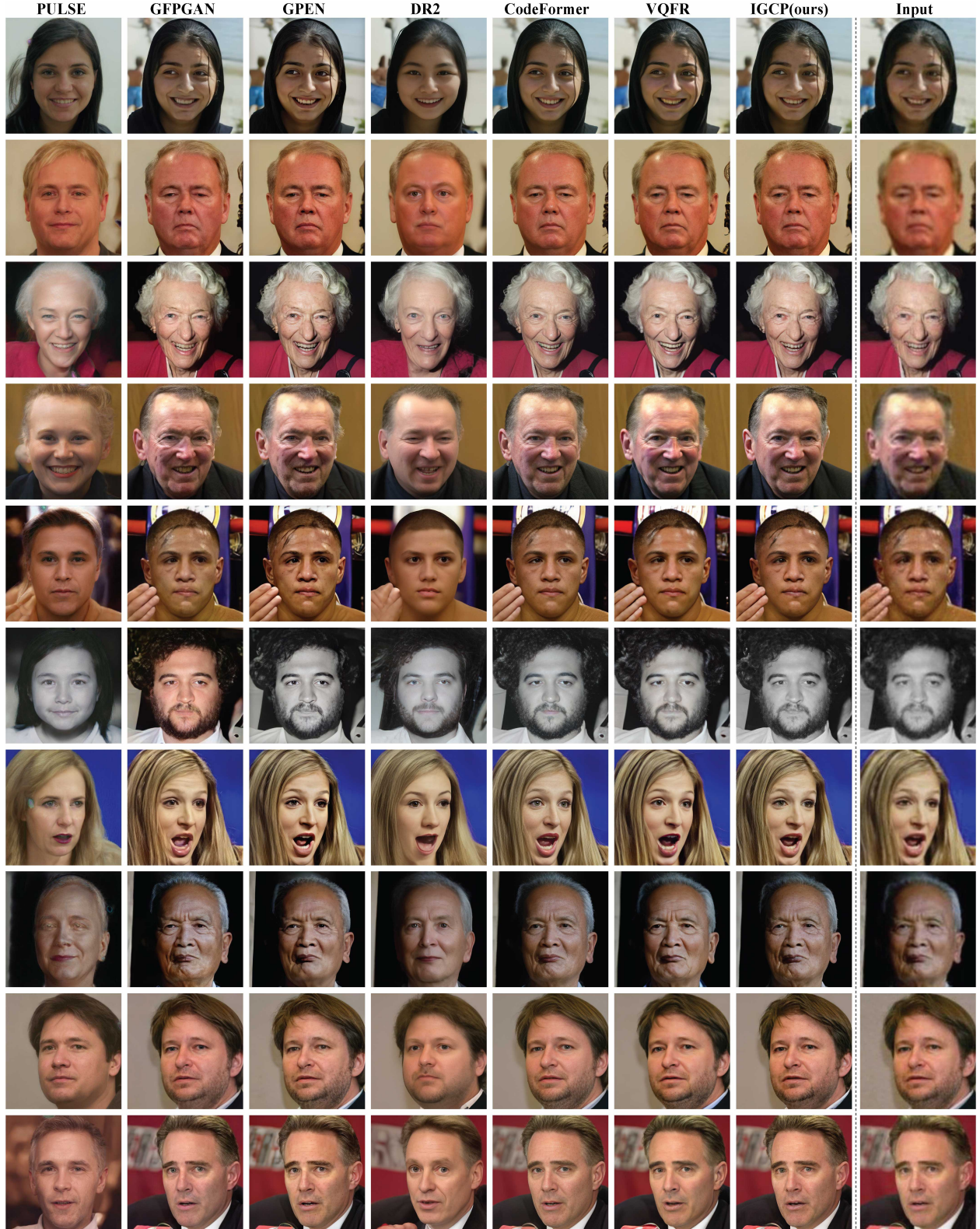


Figure 4.13: Qualitative comparison of IGCP with SOTA models on **LFW-Test** dataset [16] (mild degradation).

Figure 4.13 provides a qualitative comparison between IGCP and other SOTA models using the lfw-Test dataset, which is marked by mild degradation. Due to the subtle nature of the degradation, most SOTA models typically generate restorations with high quality and fidelity. However, the IGCP’s unique strength comes to the fore in its ability to preserve and reflect nuanced facial expressions and gestures without overly normalizing them. An illustrative example of this capability can be observed in the second row, focusing on the man’s mouth. The right side of his mouth is distinctively downturned, lending him a specific expression. While other models seem to slightly neutralize or smoothen this aspect, IGCP retains the unique shape of the mouth. This precision enables IGCP to reflect gestures and emotions with a natural touch.

4.4.3 Human Evaluation.

Given the limitations of quantitative metrics and their often weak correlation with human visual perception of image quality, we conducted qualitative assessments through two user studies. These studies aimed to provide a subjective assessment of the effectiveness of our blind face restoration method in comparison to five SOTA models: GFPGAN [16], GPEN [17], DR2 [1], Codeformer [18], and VQFR [73]. For the first study, we randomly selected 30 LQ-HQ image pairs from the CelebA-IntenseTest with the HQ images serving as references, while the LQ counterparts were restored using each of the aforementioned models. 28 participants compared the reference with the restored images from each model and ranked them on a scale of 1 (most faithful) to 6 (least faithful) based on similarity.

The second study took a similar approach but used 15 randomly chosen LQ images from the WebPhoto-Test [16] and WIDER-Test [18] datasets. Given the absence of HQ ground-truth images in these datasets, LQ images served as the reference.

Table 4.4: Average ranks of each method in human evaluation ¹². GT means ground-truth.

Study Method	GT as reference (1st)		LQ as reference (2nd)	
	Mean	STD	Mean	STD
GFPGAN	5.15	0.96	4.10	1.64
GPEN	3.91	1.49	3.14	1.48
DR2	4.49	1.31	5.34	1.24
CodeFormer	3.10	1.45	2.92	1.31
VQFR	2.64	1.24	3.05	1.39
IGCP	1.70	0.99	2.46	1.41

¹² PULSE is omitted due to its notably inferior performance.

Table 4.4 tabulates the average rank and standard deviation (STD) of the ranks for each method across both studies. IGCP consistently performed the best, achieving the lowest rank. The small STD values suggest a broad consensus among evaluators, reinforcing the reliability of the results.

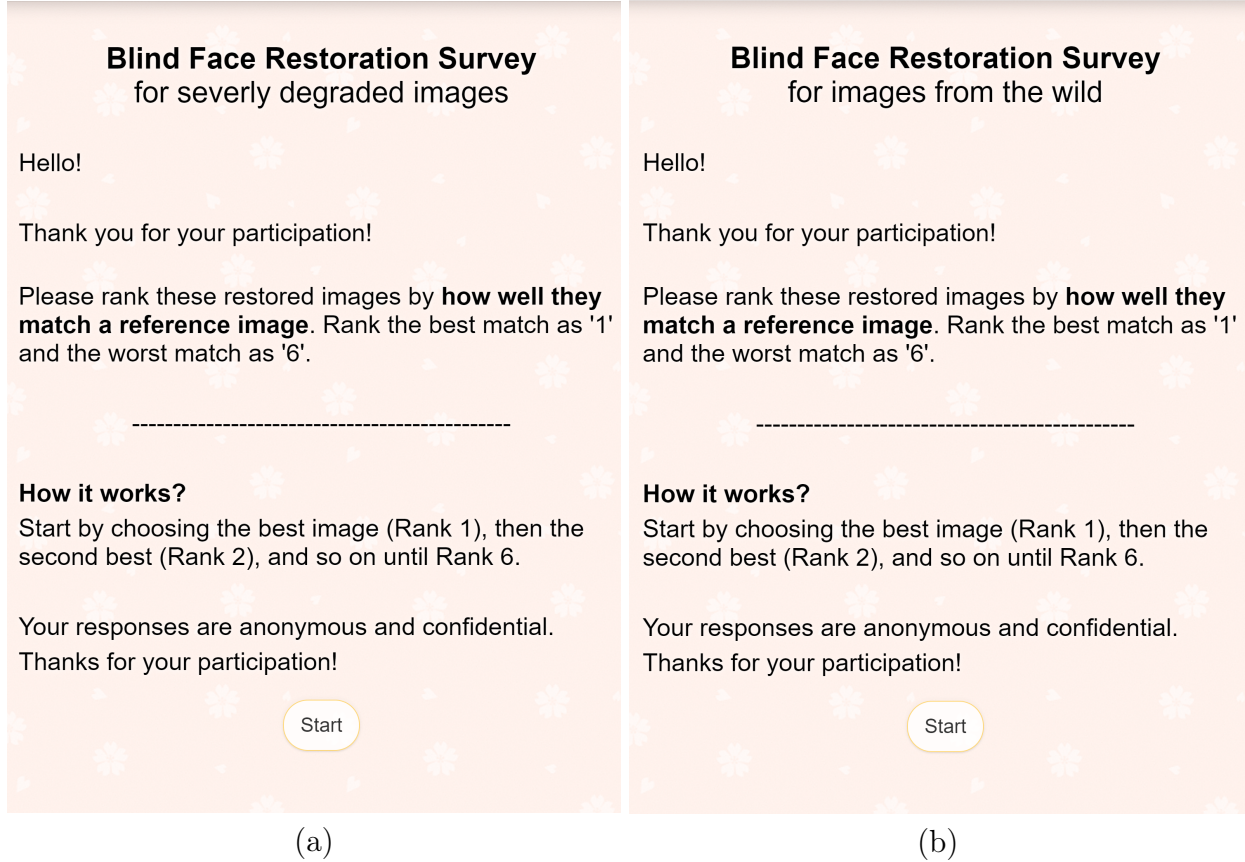


Figure 4.14: Introduction page on the two surveys, where participant are asked to rank the given image based on how well it matches to its reference.



Figure 4.15: Results from two user studies on image restoration. The left side presents the first study where 30 HQ images from the CelebA-IntenseTest dataset were randomly and restored using six different methods: GFPGAN [16], GPEN [17], DR2 [1], Codeformer[18], VQFR [73], and IGCP. The bar chart under the images summarizes the percentage of votes each method received for ranks 1 (most faithful) to 6 (least faithful). Participants ranked the restored images based on their similarity to the reference. The right side depicts the second study, which used 15 LQ images from the WebPhoto-Test and Wider-Test datasets.

In Figure 4.15, we provide two examples of images shown to the user, one from each study. The left-hand side is an example of the first study where 30 HQ images from the CelebA-HQ dataset were degraded and then restored using six different methods: GFPGAN [16], GPEN [17], DR2 [1], Codeformer[18], VQFR [73], and our proposed model denoted as IGCP. The reference image (presented at the top) in the first study is the ground-truth HQ

image. The right-hand side depicts an example from the second study, which used 15 LQ images from the WebPhoto-Test and Wider-Test datasets, where no ground-truth image is provided. The LQ image (shown on the top) is presented to the user as the reference image in this second study. In both studies, the participants ranked the restored images based on their similarity to the reference. In Figure 4.15, we also show a bar chart for each study, summarizing the percentage of votes each method received for ranks 1 (most faithful) to 6 (least faithful).

4.5 Ablation Study

4.5.1 GPR: Assessing Identity to Style Integration.

We assessed the merit of integrating identity with style in the GPR network through three model variations: 1. GPR/wo/i: trained with no identity loss. 2. GPR/wo/i_s: incorporated only the identity-in-image loss ignoring the proposed identity-in-style loss, and 3. GPR: integrated both the identity-in-image and identity-in-style losses. Each configuration was tested on 200 degraded CelebA-HQ images across 10 identities (20 images for each). The t-SNE [117] visualizations of style vectors from these variants are shown in Figure 4.16.

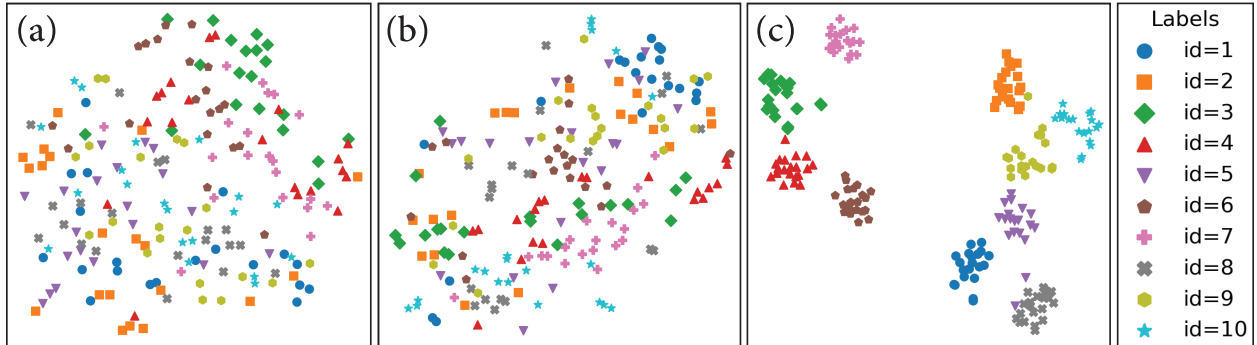


Figure 4.16: Comparison of the style vectors generated by three variants of IGCP: (a) GPR/wo/i, which does not use any identity loss, (b) GPR/wo/i_s, which uses only identity-in-image loss, and (c) GPR, which uses both identity-in-image and identity-in-style losses. We use t-SNE [117] to visualize the embeddings of the style vectors in a two-dimensional space. We observe that the style vectors are clustered by identity when we use the identity-in-style loss ($\mathcal{L}_s$), indicating that it effectively preserves the identity information within the style vectors.

as shown in Figure 4.16, The GPR model is the only one that exhibits clustering by identity, highlighting the effectiveness of our identity-to-style integration approach. This is supported by the visual comparison provided in Figure 4.17, as the image from the GPR model closely mirrors the ground-truth and achieves a superior IDS score.

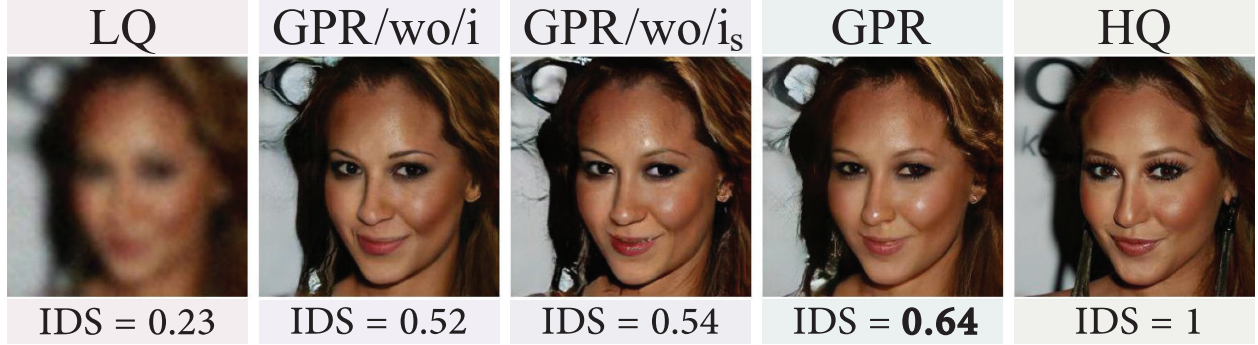


Figure 4.17: Comparisons of different GPR configurations.

To evaluate the quality of our identity embeddings in style vectors, we compute three key metrics: intra-class and inter-class distances. These distances are derived from the identity vectors of generated images:

- **Intra-Class Distance:** This metric computes the average dissimilarity of identity vectors extracted from images within the same identity class, indicating the model’s consistency in preserving unique identity features. A lower distance signifies a superior ability to capture the distinct characteristics of the same identity across various images.
- **Inter-Class Distance:** This measures the mean dissimilarity between identity vectors extracted from images belonging to different identity classes, with a higher distance being preferable. It reflects the model’s effectiveness in distinguishing between disparate identities.
- **Separability Index:** Defined as the ratio of intra-class to inter-class distance, this index quantifies the distinguishability of identity classes. A value approaching 0 denotes efficient separation between identities, highlighting the model’s proficiency in not only differentiating between various identities but also in maintaining consistency within images of individual identities.

Both intra-class and inter-class distances are calculated based on the cosine similarity between feature vectors. Cosine similarity ranges from -1 to 1, where 1 represents identical orientation, 0 indicates orthogonality, and -1 implies the opposite directions. For our purposes, we transform cosine similarity into a distance measure by subtracting it from 1, resulting in

a range of 0 to 2, where 0 indicates no distance (maximum similarity) and 2 indicates the maximum possible distance.

Table 4.5 summarizes the intra-class and inter-class distances for the ten different identity classes in Figure 4.16. These metrics confirm the effectiveness of our identity-in-style integration approach.

Table 4.5: Comparative analysis of intra-class and inter-class distances of identity vectors from 10 identity classes, each represented by 20 examples. These vectors are derived from images restored by the GPR network under various configurations, alongside comparisons to ground-truth HQ images.

IES Configuration	Average IDS $\uparrow$	Intra-class Distance $\downarrow$	Inter-class Distance $\uparrow$	Separability Index $\downarrow$
Input LQ	0.36	0.63	0.81	0.78
IES/wo/i	0.52	0.65	0.89	0.73
IES/wo/is	0.57	0.62	0.91	0.68
IES	0.61	0.60	0.95	0.63
Ground-truth HQ	1.00	0.37	0.99	0.37

4.5.2 CPR: Significance of FTM Module.

To explore the impact of the FTM module, we used the output image from the GPR module, denoted as GPR, and fed it to several CPR configurations: 1. CPR model without the FTM module containing only a pre-trained VQAE with no finetuning denoted as CPR/wo/ft. 2. CPR module without the FTM but with VQAE finetuning, referred to as CPR/wo/FTM. 3. CPR: The final CPR module that involves VQAE finetuning with the aid of the proposed FTM. The images produced by these configurations are illustrated in Figure 4.18. Our findings suggest that the inclusion of the FTM module within the CPR model along with finetuning the VGAE network yields marked enhancements both in enhancing image quality and maintaining the identity of the subject in the input image.

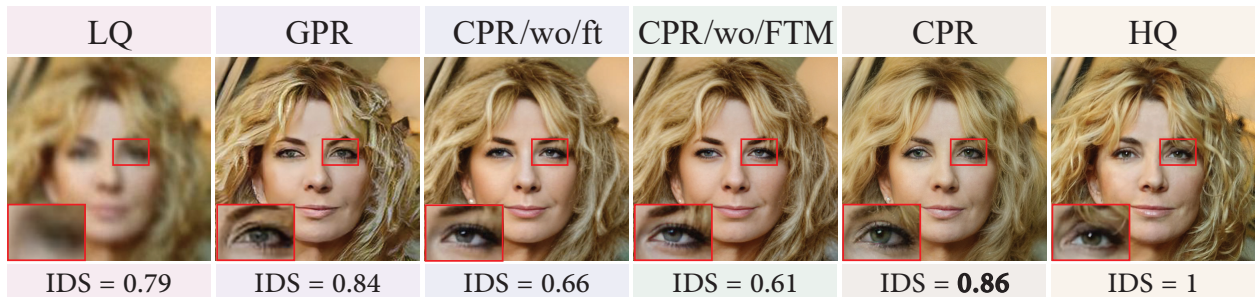


Figure 4.18: Comparisons of different CPR configurations.

4.6 Extension to other applications

The proposed IGCP framework is versatile and can be easily adapted to various face manipulation tasks beyond restoration. We have expanded its application to tasks such as super-resolution, image coloring, and inpainting. Preliminary results suggest our framework has the potential to achieve high performance in these domains.

4.6.1 Inpainting

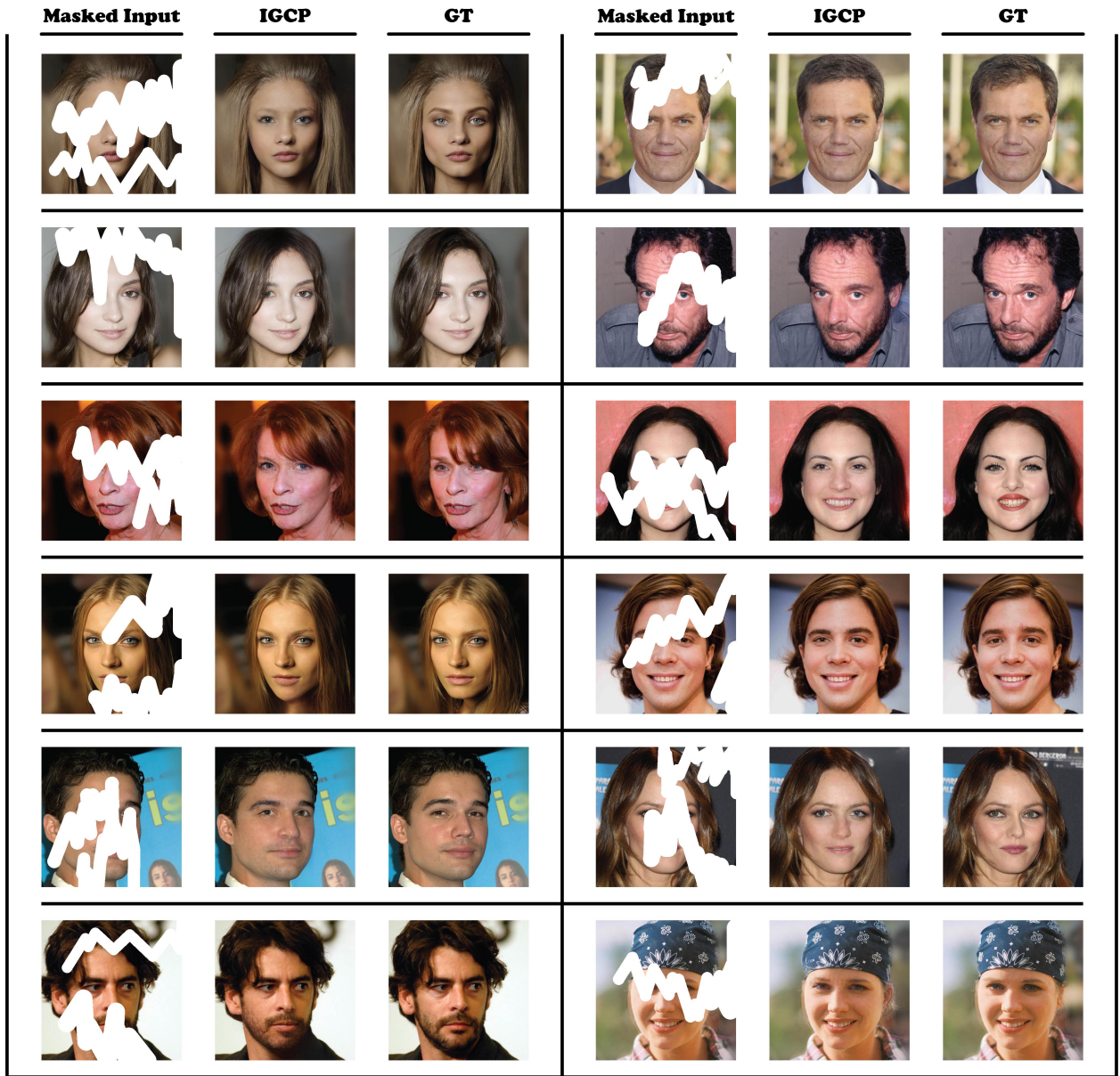


Figure 4.19: IGCP face inpainting results.

The IGCP model can be adapted for face inpainting, where it demonstrates remarkable performance. To construct the training pairs, we utilized a publicly accessible script [17] to generate masked faces by randomly drawing irregular polyline masks. The corresponding results are illustrated in Figure 4.19. Our IGCP approach produces realistic facial images of high quality, without strokes and artifacts.

4.6.2 Super-resolution

Figure 4.20 demonstrates the application of the IGCP model in super-resolution tasks. Specifically, it highlights the model’s capability to reconstruct an image from its Low Resolution (LR) counterpart that was downsampled by a factor of 16. The resultant images indicate the model’s efficacy in restoring essential features and details, which emphasizes the viability of IGCP in high-ratio super-resolution contexts.

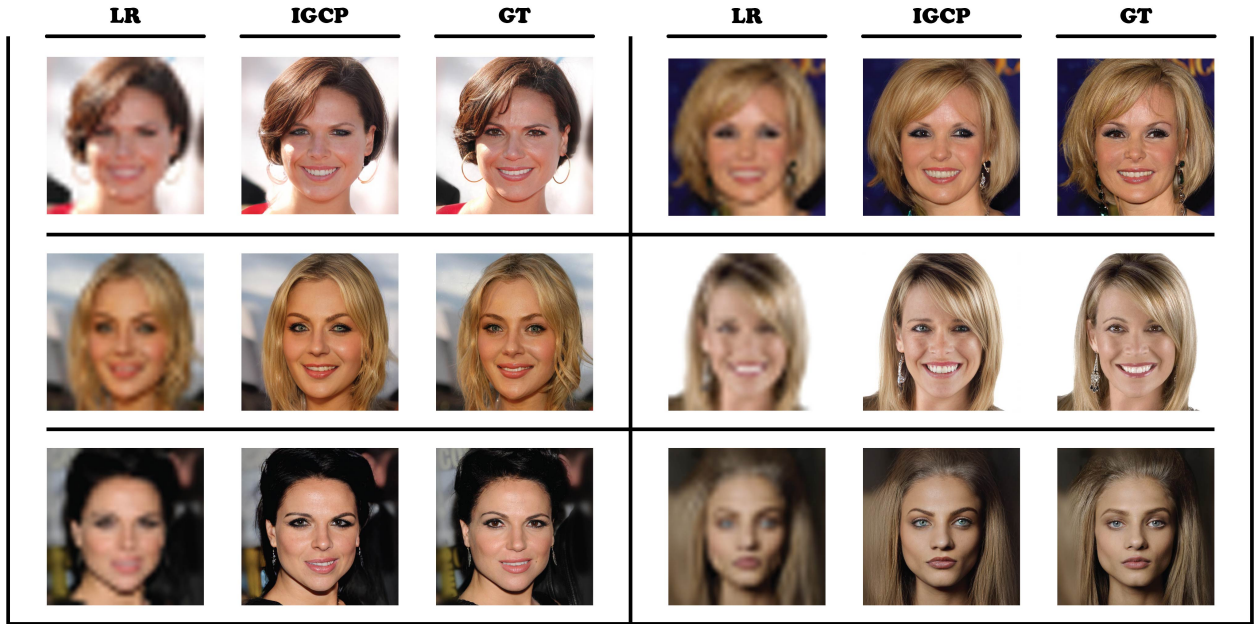


Figure 4.20: IGCP super resolution results (I): the Low Resolution (LR) image is downsampled by a factor of 16.

Figure 4.21 further attests to the adaptability and performance of the IGCP model in super-resolution endeavors. Here, the challenge is even more pronounced, as the model is tasked with the super-resolution of images downsampled by a factor of 32. Despite the substantial loss of detail inherent to such a severe reduction, IGCP produces commendably detailed restorations, underlining its potential for applications even in extreme super-resolution scenarios.

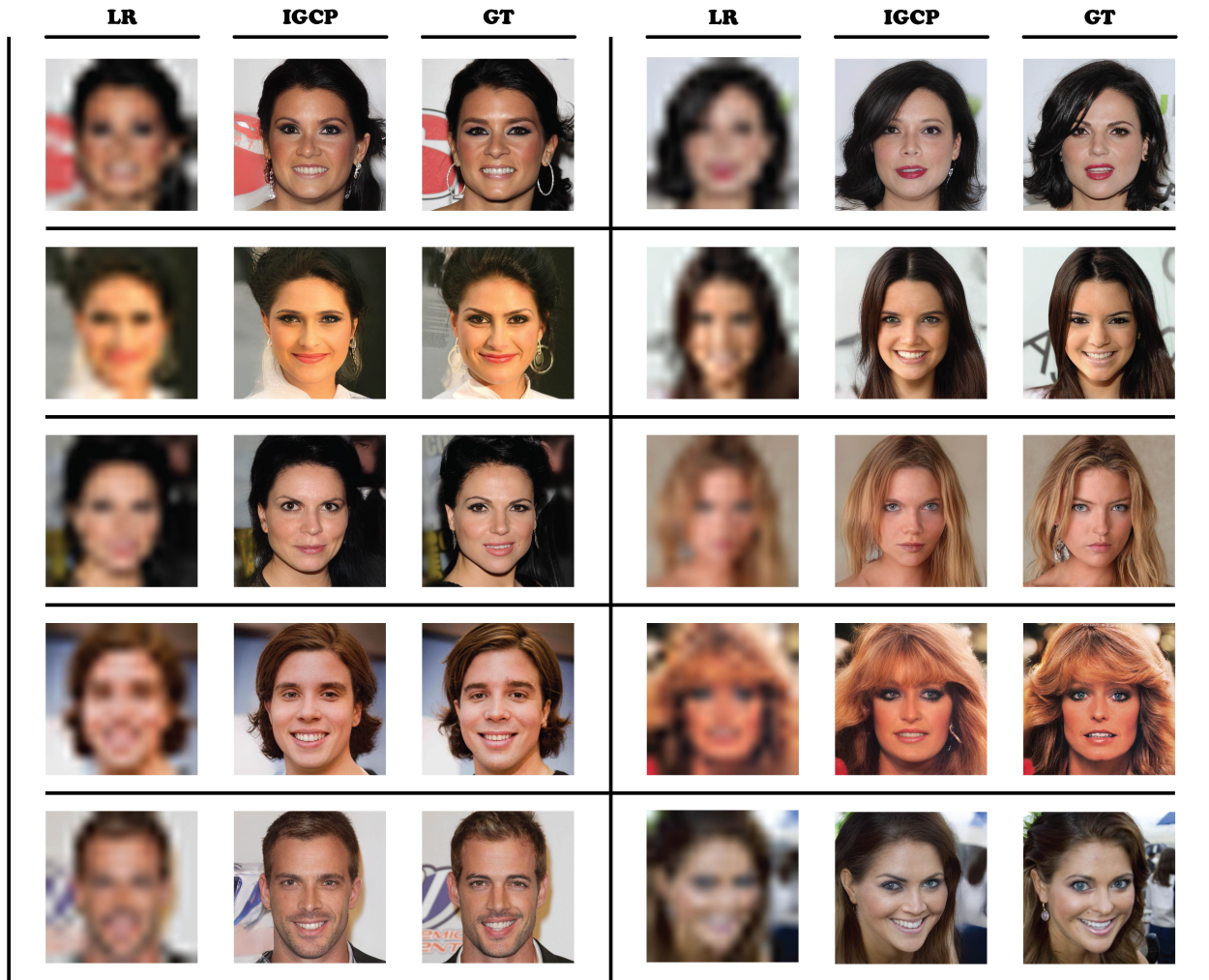


Figure 4.21: IGCP super resolution results (II): the Low Resolution (LR) image is downsampled by a factor of 32.

4.6.3 Colorization

We leveraged the capabilities of the IGCP model to address the intricate task of colorizing grayscale images. The outcomes of our experiments with the IGCP model in this domain have been promising. As illustrated in the Figure 4.22, the model successfully transforms grayscale images into their colorized counterparts, rendering them with commendable accuracy.

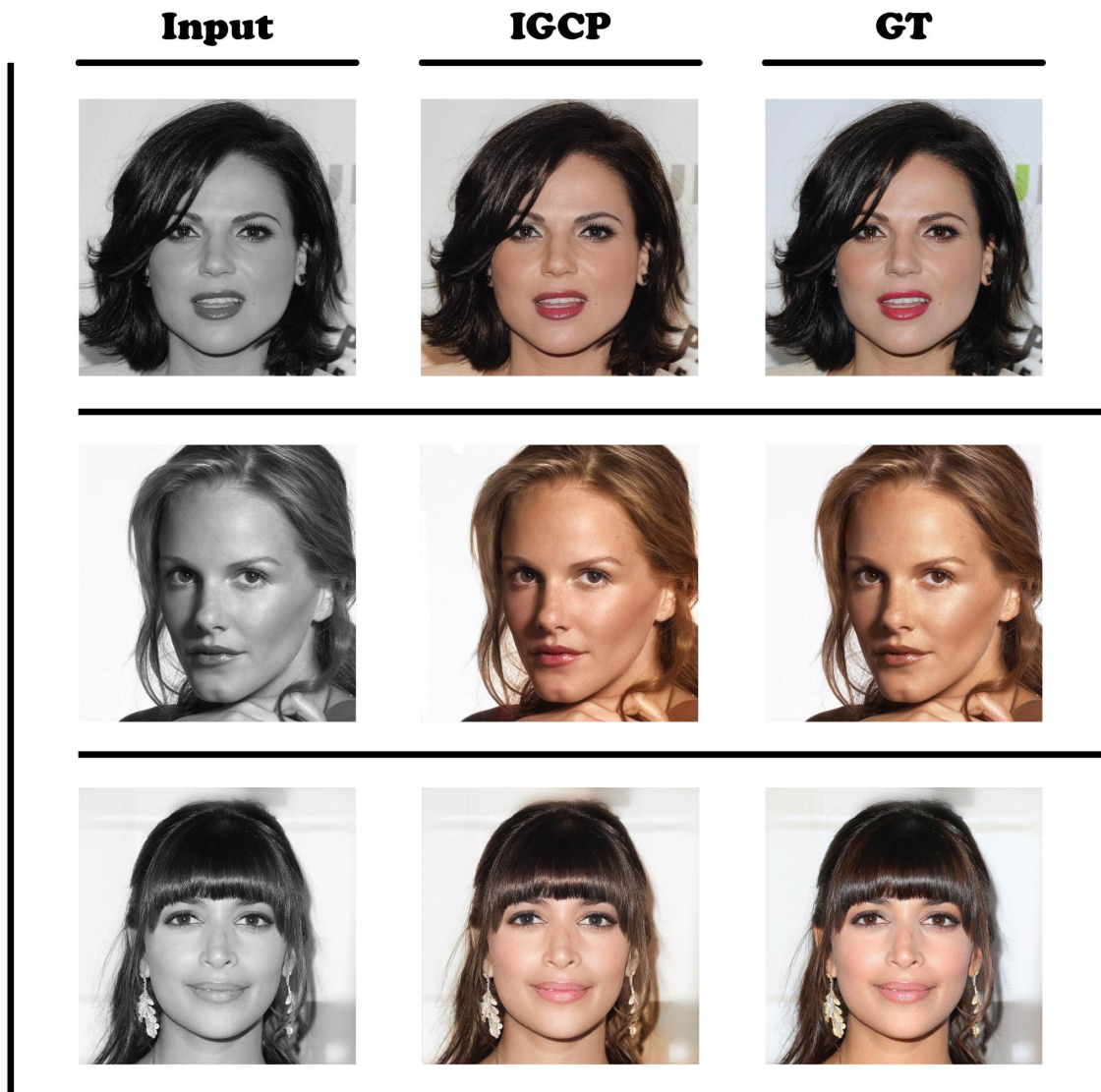


Figure 4.22: IGCP colorization results on gray-scale images from CelebA-HQ dataset.

Chapter 5

Conclusion and future work

In this thesis, we present a novel framework for blind face restoration (BFR) that combines an Identity-embedded GAN Prior Restorer (GPR) module with a Codebook Prior Restorer (CPR). This unique combination of GAN priors and codebook priors allows our BFR model to effectively handle varying levels of degradation in face images.

In the GPR model, we propose integrating a U-shaped autoencoder with a generative prior network. By using an autoencoder, we successfully eliminate initial distortions from both style and spatial features. Additionally, we enhance fidelity by directly merging the autoencoder features with the generative features in successive generative blocks. We introduce a novel identity-in-style loss function that infuses identity information into the style vector, further improving the quality of the restored faces.

For the CPR model, we employ a quantized autoencoder to rectify any unresolved corruptions in the output of the GPR model. Through fine-tuning of the quantized autoencoder and incorporating a feature-transforming module, we amplify the conveyance of fidelity information from encoder features. This approach significantly enhances fidelity while maintaining overall quality.

Our extensive experiments demonstrate that the innovative design components in our framework collectively contribute to superior performance in terms of fidelity, quality, and identity preservation. We have also identified an intriguing application of our framework for attribute-based face hallucination, such as gender and hair color, which we plan to explore in future work.

In addition to the advancements we have made in blind face restoration, we envision an intriguing application of our framework in the realm of conditional blind face restoration. Specifically, we aim to extend our framework to enable the generation of conditioned outputs based on facial attributes such as gender and hair color.

At present, our framework primarily focuses on blind face restoration without considering specific facial attributes. However, by incorporating a conditional generation component, we can enhance the versatility and practicality of our framework. This extension would allow users to specify desired facial attributes, and the model would generate restored faces that align with those attributes.

To achieve this, we would need to incorporate attribute recognition and manipulation modules into our framework. These modules would enable the identification and manipulation of specific facial attributes, thereby guiding the restoration process according to the user's preferences.

By expanding our framework to encompass conditional face restoration, we can provide users with more control over the restoration process and offer a wider range of applications. For instance, users could generate restored faces with different gender presentations or experiment with various hair colors.

Furthermore, we aspire to explore the integration of text and natural language processing to further refine the quality of generated images. This approach will involve interpreting descriptive text or contextual information about the desired image attributes, such as eye or hair color, or specific facial expressions. Utilizing advanced language models like GPT-4¹³, we plan to process this input text, extract relevant features, and provide these as additional parameters to our image generation system. This synergy between natural language processing and image restoration holds the potential to revolutionize the way we interact with and guide image generation technologies, making them more intuitive and aligned with human-centric descriptions and requests. Such advancements will not only enhance the technical robustness of our framework but also open new avenues for creative and personalized image restoration and generation.

¹³ <https://doi.org/10.48550/arXiv.2303.08774>

Appendix A

Appendix

A.1 StyleGAN

The Style Generative Adversarial Network (StyleGAN) is an extension of the GAN architecture that introduces significant modifications to the generator model. It is capable of producing HQ, photorealistic images with a resolution up to 1024x1024 pixels. The key contribution of StyleGAN is the introduction of a novel style-based generator architecture that significantly improves the control over the generated images' style and content. This innovative approach separates high-level attributes (such as pose and identity) from low-level details (such as textures and colors) in the image generation process. This separation allows for more fine-grained control over the generated images and makes it easier to manipulate specific aspects of the images while keeping others unchanged.

In this thesis, StyleGAN serves as a fundamental component of the proposed methodology. Consequently, an in-depth examination of the two prominent versions of StyleGAN will be conducted to elucidate their respective characteristics and contributions to the field.

StyleGAN builds upon the progressive growing GAN architecture, which incrementally develops both the discriminator and generator models, starting from LR images and progressively increasing their size. However, unlike traditional GANs, StyleGAN's generator does not take a latent vector as input. Instead, it starts with a constant matrix and uses two new sources of randomness to generate synthetic images: style vector and noise layers.

The mapping network produces a style vector that controls the features of the generated image at different levels of detail. This style vector is integrated into the generator model at specific points through a new layer called adaptive instance normalization (AdaIN). Additionally, stochastic variation is introduced through noise layers added at specific points

in the generator model. These noise layers are applied to entire feature maps, enabling the model to learn styles in a fine-grained, per-pixel manner.

StyleGAN adopts the baseline progressive GAN architecture with some modifications to the generator part. The discriminator architecture remains largely similar to the baseline progressive GAN. Some of the architectural differences include:

Baseline Progressive Growing GANs: StyleGAN employs the progressive GAN structure, wherein the size of the generated image increases progressively from low resolution (4x4) to high resolution (1024x1024). New layers are added to both the generator and discriminator models to accommodate larger resolutions.

Bi-linear Sampling: In both the generator and discriminator, bi-linear sampling is used instead of the nearest neighbor up/down sampling used in previous baseline progressive GAN architectures.

Mapping Network and Style Network: The mapping network produces an intermediate vector from the latent input vector, with each element controlling various visual features. This mapping replaces the direct application of the latent vector to the input layer.

Elimination of Traditional (Latent) Input: The authors of StyleGAN found that the photo-generation characteristics are examined by the intermediate vector (w) and AdaIN. As a result, they replaced the initial input with a constant matrix of dimensions 4x4x512 to improve the network’s performance.

Noise Addition: Gaussian noise is added to each activation map before the AdaIN operation. A separate noise sample is calculated for each layer based on its scaling factors.

Merging Regularization: The StyleGAN uses the intermediate vector at various levels of the synthesis network, enabling the network to learn the relationship between different levels of detail.

A.1.1 StyleGAN v1

The primary contribution proposed in StyleGAN v1 lies in the innovative architecture of the generator, which proposes numerous enhancements over the conventional approaches. the StyleGAN generator, consists of two networks mapping and synthesis.

A.1.1.1 Mapping Network

The mapping network in StyleGAN (see fig. [A.1](#)) is an essential component that introduces an intermediate latent space, known as the W space. Unlike the traditional

approach, where a vector from the Z space is directly fed to the generator, the W space is formed by mapping Z space vectors through an 8-layer Multilayer Perceptron (MLP). The primary purpose of introducing the W space is to achieve better disentanglement. When the features are disentangled, it means that each dimension of the W space vector holds unique information about the image, such as pose, facial expression, or hair color. This organized representation enables more precise control over the image synthesis process, making it possible to modify specific characteristics while keeping other aspects of the image unchanged leading to generation of more realistic and diverse images, as well as perform intuitive and effective mixing and interpolation operations between different images.

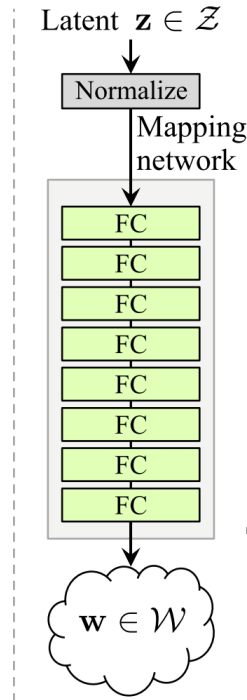


Figure A.1: Mapping network in StyleGAN v1, depicting the transformation of a 512-dimensional input vector drawn from a normal distribution into a 512-dimensional output vector ($\mathbf{w}$) using an 8-layer Multilayer Perceptron (MLP). Each dimension of the output vector $\mathbf{w}$ contains distinct information related to various image attributes.

To determine whether the W space truly presents less entanglement than the Z space, the authors propose two new metrics: Perceptual path length, which measures the degree of changes in the image during interpolation, and Linear separability, which calculates the ease of separating latent space points corresponding to two image classes using a hyperplane.

The authors demonstrate that the W space, combined with a style-based generator architecture, yields the best Frechet Inception Distance (FID) score, perceptual path length, and separability. They also show that adding a mapping network to a traditional generator results in a loss of separability but a better FID, emphasizing the strengths of the W space.

Table A.1: Performance analysis of various mapping network depths on the FFHQ dataset. As demonstrated that FID, separability, and path length are all positively affected by the presence of a mapping network, for both style-based and traditional generator architectures. Additionally, deeper mapping networks tend to outperform shallower ones, further emphasizing the benefits of incorporating a mapping network.

Method	FID	Path length		Separa- bility
		full	end	
B Traditional 0 $\mathcal{Z}$	5.25	412.0	415.3	10.78
Traditional 8 $\mathcal{Z}$	4.87	896.2	902.0	170.29
Traditional 8 $\mathcal{W}$	4.87	324.5	212.2	6.52
Style-based 0 $\mathcal{Z}$	5.06	283.5	285.5	9.88
Style-based 1 $\mathcal{W}$	4.60	219.9	209.4	6.81
Style-based 2 $\mathcal{W}$	4.43	217.8	199.9	6.25
F Style-based 8 $\mathcal{W}$	4.40	234.0	195.9	3.79

A.1.1.2 Synthesis Network

The synthesis network in StyleGAN v1 plays a pivotal role in generating the final image output based on the styles provided by the mapping network and the affine transformations. The key components of the synthesis network are as follows:

- **Affine Transform:** The learned affine transformation is essential for converting the w vectors into styles, which are then fed into the synthesis network. This transformation, referred to as "A" in the original paper, allows each layer of the synthesis network to receive one style vector, enabling cross-layer style control and improving the model's performance.
- **AdaIN (Adaptive Instance Normalization):** This technique comprises normalization followed by modulation and was initially introduced in a style transfer model. It facilitates the control of style features in the generated image.
- **Noise – B:** The primary objective of introducing noise into the synthesis network is to create stochastic variations in the output image. This concept can be compared to the imperceptible changes in the positions of stars in the sky. Similarly, small features like hair, freckles, and skin pores in human faces exhibit stochastic behavior. This randomization is achieved by adding per-pixel noise after each convolution.

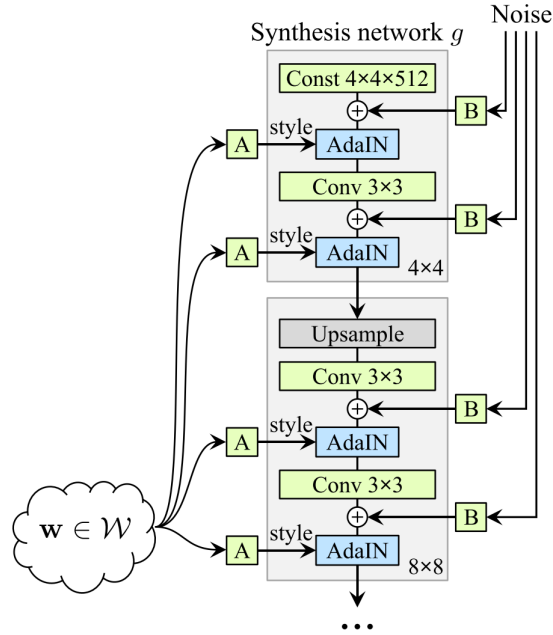


Figure A.2: StyleGAN v1 synthesis network. The synthesis network (g) is comprised of 18 layers, with two layers allocated for each resolution ranging from 4x4 to 1024x1024 pixels. The latent vector w in the latent space $\mathcal{W}$, controls the generator using Adaptive Instance Normalization (AdaIN) at every convolutional layer. Following each convolution, Gaussian noise is incorporated prior to computing the nonlinearity. In this context, "A" denotes a learned affine transformation, while "B" signifies the application of learned per-channel scaling factors to the noise input. Finally, a separate 1x1 convolution is employed to convert the output of the last layer to RGB format.

A.1.1.3 StyleGAN v1 Issues

StyleGAN v1 has a few notable issues that were addressed in the subsequent version, StyleGAN2. Some of these issues include:

- **Blob-like artifacts:** In StyleGAN-generated images, there are observable blob-like artifacts that resemble water droplets. These artifacts become more prominent when examining the intermediate feature maps within the generator network. The issue seems to begin at a resolution of 64x64 in all feature maps and worsens at higher resolutions.



Figure A.3: blob-like artifacts in In StyleGAN v1. In StyleGAN v1, blob-like artifacts resembling water droplets sometimes appear in the generated images. Although these artifacts are not always visible in the final output, the problem is always there when examining the activations within the generator network. The presence of these artifacts can be traced back to the use of instance normalization in the Adaptive Instance Normalization (AdaIN) operation.

- The progressive growing utilized in StyleGAN v1, involves gradually increasing the resolution of the images generated by the GAN, starting from LR images and progressively adding layers to the generator and discriminator models to accommodate higher resolutions. While progressive growing has been effective in generating HQ images, it also introduced some challenges. One of the main issues with progressive growing is that it can lead to some artifacts and inconsistencies in the generated images, especially as the resolution increases. This is because, as new layers are added to the models, the network has to adapt to the changes in architecture, which can result in instability during training.

To address these issues, StyleGAN2 proposes an alternative design that eliminates the need for progressive growing. Instead, it introduces a number of architectural improvements, such as weight demodulation and revised AdaIN, to stabilize HR training without the need for the progressive growing technique. These changes help to overcome the limitations of progressive growing and result in higher-quality images with fewer artifacts and inconsistencies.



Figure A.4: Phase artifacts resulting from progressive growing in StyleGAN v1. In this example, the teeth remain aligned with the camera (indicated by the blue line), despite the generated faces turning to the right. This demonstrates that progressive growing can lead to unnatural and misaligned facial features in the generated images.

A.1.2 StyleGAN v2

StyleGAN v2 is an improved version of StyleGAN v1, featuring several key differences that result in better image quality, fewer artifacts, and more stable training. The main differences between StyleGAN v1 and StyleGAN v2 are:

A.1.2.1 Weight Demodulation

Supported by the experimental findings (see fig.), the following modifications made in StyleGAN2:

1. Simplification of the initial constant processing.
2. Exclusion of the mean during feature normalization.
3. Relocation of the noise module outside the style module, enhancing the overall structure.

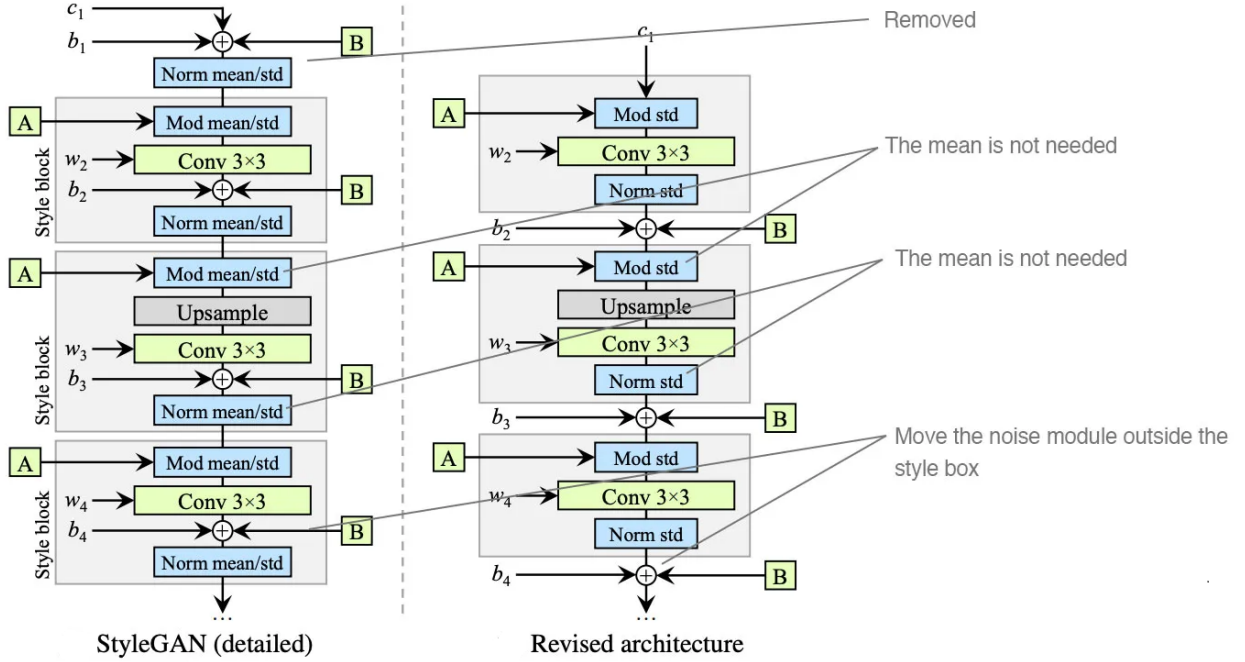


Figure A.5: Revised architecture in StyleGAN2. This illustration showcases the modifications made to the original architecture, including the removal of redundant operations at the beginning, repositioning of the addition of b and B outside the active area of a style, and adjusting the standard deviation per feature map.

4. Incorporation of the Mod, Demod operations: Other than the previous architectural modifications, the following changes have been applied in StyleGAN v2:

The modulation (mod std) and the subsequent convolution (Conv 3x3) can be merged by scaling the convolution weights, which is implemented as Mod above. This alteration does not affect the model's overall design.

$$w'_{ijk} = s_i \cdot w_{ijk}$$

where i is the input feature map i .

Next, the weight is normalized using the Demod operation:

$$\sigma_j = \sqrt{\sum_{i,k} w'_{ijk}{}^2}$$

The updated weight is normalized as follows:

$$w''_{ijk} = \frac{w'_{ijk}}{\sqrt{\sum_{i,k} w'_{ijk}{}^2 + \epsilon}}$$

A small value of ϵ is added to prevent numerical instability. Although this process is not mathematically identical to instance normalization, it normalizes the output feature map to maintain a standard unit standard deviation, thus achieving objectives similar to other normalization methods. In particular, it enhances the stability of the training process. Experimental results indicate that this approach successfully eliminates blob-like artifacts.

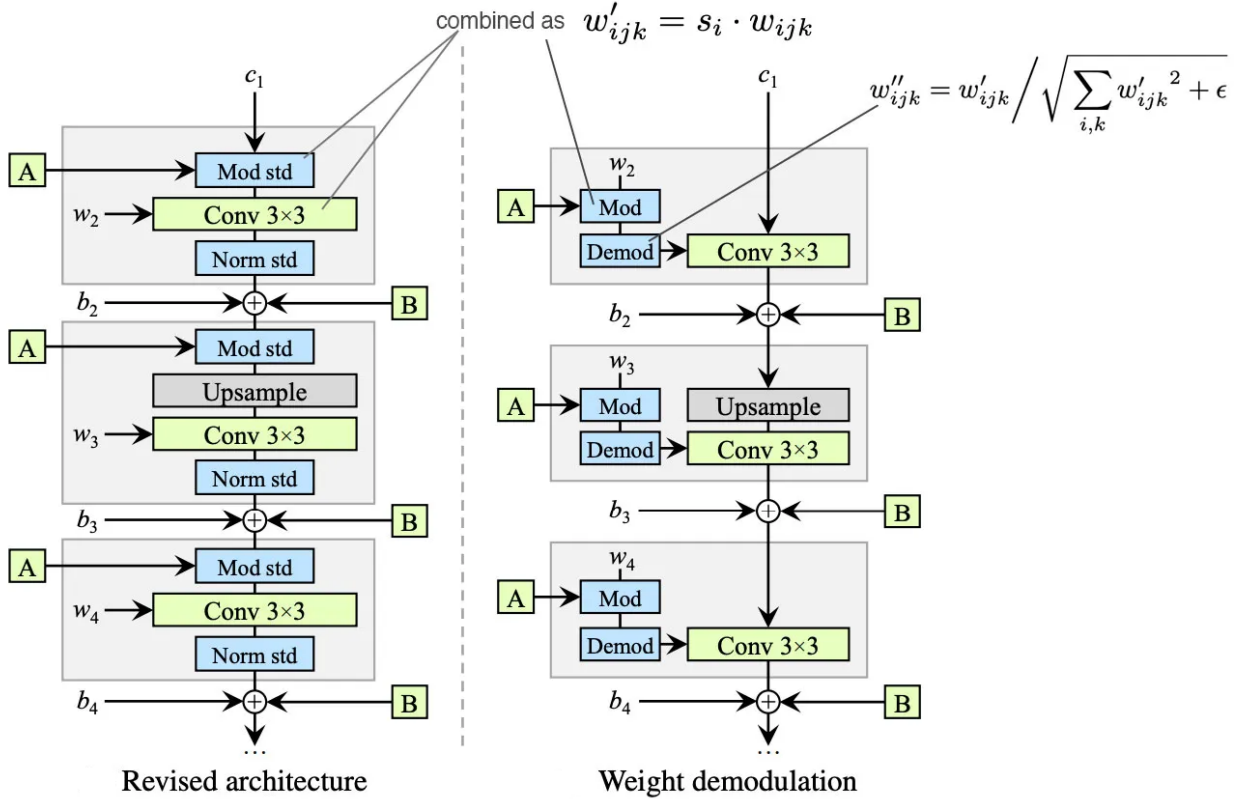


Figure A.6: Replacement of instance normalization with a “demodulation” operation being applied to the weights associated with each convolution layer, as explained in [76].

A.1.2.2 Lazy Regularization

In StyleGAN v2, the authors introduce an improvement called lazy regularization, which stems from the observation that the model’s performance remains unaffected when regularization costs are calculated less frequently. In the context of StyleGAN, R_1 regularization is applied when training with the FFHQ dataset. However, it has been discovered that ignoring the regularization cost most of the time during cost calculation does not negatively impact the model’s performance. In fact, incorporating the regularization cost only once every 16 mini-batches maintains the model’s effectiveness while simultaneously reducing computational costs. This innovative regularization technique, known as lazy regularization, enables more efficient training without compromising the model’s performance.

A.1.2.3 Path length regularization

In StyleGAN v2, path length regularization serves to improve the disentanglement and interpretability of the latent space while also acting as a performance measure for GANs. A potential concern arises when there is significant variation in path distances between different segments along the interpolation path. Ideally, linearly interpolated points should maintain consistent image distances between consecutive points. This consistency implies that equal displacements in the latent space should correspond to equivalent magnitudes of change in the image space, irrespective of the latent factor value.

To address this issue, an additional cost is incorporated when the change in the image space diverges from the ideal expected displacement. The change is calculated using the gradient available for backpropagation, and the expected displacement is approximated by the running average observed thus far. The path length regularization term is integrated into the generator’s loss function, encouraging the model to generate small and consistent changes in the output image for slight adjustments in the input latent code.

Consequently, the latent space in StyleGAN v2 becomes smoother and more interpretable, facilitating more controlled manipulations of the generated images. Path length regularization plays a crucial role in enhancing the quality and stability of StyleGAN v2 compared to its predecessor, StyleGAN v1.

A.1.2.4 Larger Network Architecture

In order to perform an in-depth analysis to investigate the influence of fine resolution layers on image generation, the authors of StyleGAN v2 measured the contribution of various layers to the variance of the generated images. The diagram on the left in Fig. A.7 illustrates the contribution of each layer, with the horizontal axis representing the progress of the training process. Initially, the lower resolution layers are expected to dominate. However, as the training progresses, the HR layers should gradually become more prominent in their contribution.

Despite this expectation, the contributions from the HR layers, specifically the 1024×1024 layers, were not as substantial as anticipated. It was hypothesized that the capacity of these layers was insufficient. To address this, the number of feature maps in the HR layers was doubled, which led to a significant improvement in their impact, as illustrated by the diagram on the right in Fig. A.7.

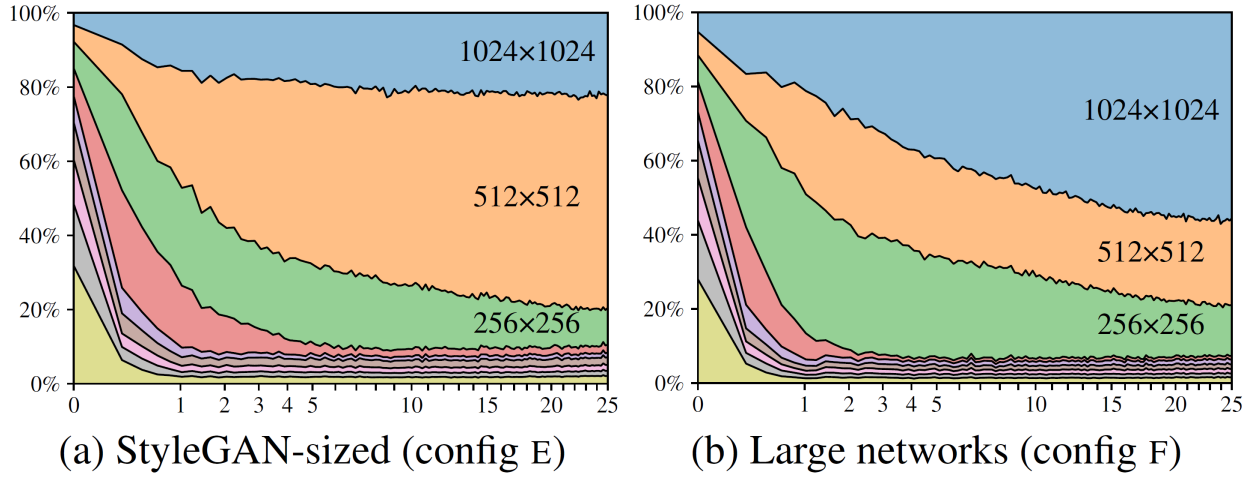


Figure A.7: StyleGAN v2 network size: The vertical axis in the diagrams depicts the relative standard deviations of different resolutions, while the horizontal axis corresponds to the training progress, measured in millions of training images shown to the discriminator. It is evident that, initially, the network focuses on LR images and progressively shifts its focus to larger resolutions as training advances. In the original configuration with the StyleGAN network size (a), the generator essentially outputs a 512×512 image with minor sharpening for 1024×1024 , whereas in the modified configuration (b), the larger network concentrates more on HR details.

Bibliography

- [1] Z. Wang, Z. Zhang, X. Zhang, H. Zheng, M. Zhou, Y. Zhang, and Y. Wang, “Dr2: Diffusion-based robust degradation remover for blind face restoration,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1704–1713, 2023. [ix](#), [2](#), [27](#), [86](#), [96](#), [98](#)
- [2] E. Luo, S. H. Chan, and T. Q. Nguyen, “Adaptive image denoising by mixture adaptation,” *IEEE transactions on image processing*, vol. 25, no. 10, pp. 4489–4503, 2016. [1](#)
- [3] J. Zeng, X. Qiu, and S. Shi, “Image processing effects on the deep face recognition system,” *Math. Biosci. Eng*, vol. 18, no. 2, pp. 1187–1200, 2021. [1](#)
- [4] S. Parameswaran, E. Luo, and T. Q. Nguyen, “Patch matching for image denoising using neighborhood-based collaborative filtering,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 2, pp. 392–401, 2016. [1](#)
- [5] Z. Shen, W.-S. Lai, T. Xu, J. Kautz, and M.-H. Yang, “Exploiting semantics for face image deblurring,” *International Journal of Computer Vision*, vol. 128, pp. 1829–1846, 2020. [1](#)
- [6] H. Zhang, C. Shi, X. Zhang, L. Wu, X. Li, J. Peng, X. Wu, and J. Lv, “Multi-scale progressive blind face deblurring,” *Complex & Intelligent Systems*, vol. 9, no. 2, pp. 1439–1453, 2023. [1](#)
- [7] M. Long, X. Yu, S. Cong, W. Zoujian, D. Jiangbin, and Z. Jiayao, “Face image deblurring with feature correction and fusion,” *The Visual Computer*, pp. 1–15, 2023. [1](#)
- [8] C. Liu, H.-Y. Shum, and W. Freeman, “Face hallucination: Theory and practice,” *International Journal of Computer Vision*, vol. 75, pp. 115–134, 07 2007. [2](#)

- [9] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, “Pulse: Self-supervised photo upsampling via latent space exploration of generative models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2437–2445, 2020. [2](#), [34](#), [86](#)
- [10] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015. [2](#), [17](#)
- [11] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1646–1654, jun 2016. [2](#), [17](#)
- [12] C. Chen, D. Gong, H. Wang, Z. Li, and K.-Y. K. Wong, “Learning spatial attention for face super-resolution,” *IEEE Transactions on Image Processing*, vol. 30, pp. 1219–1231, 2020. [2](#), [4](#)
- [13] A. Bulat and G. Tzimiropoulos, “Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 109–117, 2018. [2](#)
- [14] K. C. Chan, X. Wang, X. Xu, J. Gu, and C. C. Loy, “Glean: Generative latent bank for large-factor image super-resolution,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14245–14254, 2021. [2](#)
- [15] L. Yang, S. Wang, S. Ma, W. Gao, C. Liu, P. Wang, and P. Ren, “Hifacegan: Face renovation via collaborative suppression and replenishment,” in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1551–1560, 2020. [2](#), [22](#)
- [16] X. Wang, Y. Li, H. Zhang, and Y. Shan, “Towards real-world blind face restoration with generative facial prior,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9168–9178, 2021. [2](#), [8](#), [13](#), [76](#), [79](#), [82](#), [83](#), [86](#), [92](#), [94](#), [95](#), [96](#), [98](#)
- [17] T. Yang, P. Ren, X. Xie, and L. Zhang, “Gan prior embedded network for blind face restoration in the wild,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 672–681, 2021. [2](#), [8](#), [39](#), [59](#), [79](#), [86](#), [96](#), [98](#), [103](#)
- [18] S. Zhou, K. C. Chan, C. Li, and C. C. Loy, “Towards robust blind face restoration with codebook lookup transformer,” in *NeurIPS*, 2022. [2](#), [4](#), [6](#), [7](#), [13](#), [46](#), [47](#), [73](#), [77](#), [79](#), [83](#), [84](#), [86](#), [92](#), [93](#), [96](#), [98](#)

- [19] Y. Gu, X. Wang, L. Xie, C. Dong, G. Li, Y. Shan, and M.-M. Cheng, “Vqfr: Blind face restoration with vector-quantized dictionary and parallel decoder,” in *European Conference on Computer Vision*, pp. 126–143, Springer, 2022. [2](#), [79](#)
- [20] M. Bengherabi, D. Laib, F. S. Lasnami, and R. Boussaha, “Boosting cross-quality face verification using blind face restoration,” *arXiv preprint arXiv:2308.07967*, 2023. [3](#)
- [21] R. A. de Oliveira, M. H. Scheeren, P. J. S. Rodrigues, A. C. Junior, and P. L. de Paula Filho, “Super-resolution face recognition: An approach using generative adversarial networks and joint-learn,” in *Optimization, Learning Algorithms and Applications: Second International Conference, OL2A 2022, Póvoa de Varzim, Portugal, October 24-25, 2022, Proceedings*, pp. 747–762, Springer, 2023. [3](#)
- [22] S. Li, Z. Liu, D. Wu, H. Huo, H. Wang, and K. Zhang, “Low-resolution face recognition based on feature-mapping face hallucination,” *Computers and Electrical Engineering*, vol. 101, p. 108136, 2022. [3](#)
- [23] M. B. Shahbakhsh and H. Hassanpour, “Empowering face recognition methods using a gan-based single image super-resolution network,” *International Journal of Engineering*, vol. 35, no. 10, pp. 1858–1866, 2022. [3](#)
- [24] Z. Wang, Z. Luo, M. Hu, D. Wu, Y. Cao, and Y. Qin, “Revisiting super-resolution for internet video streaming,” in *Proceedings of the 32nd Workshop on Network and Operating Systems Support for Digital Audio and Video*, pp. 8–14, 2022. [4](#)
- [25] A. Zhang, C. Wang, B. Han, and F. Qian, “Efficient volumetric video streaming through super resolution,” in *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications*, pp. 106–111, 2021. [4](#)
- [26] Y. Reznik and N. Barman, “Improving the performance of web-streaming by super-resolution upscaling techniques,” *Electronic Imaging*, vol. 35, pp. 351–1, 2023. [4](#)
- [27] 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. [7](#)
- [28] X. Li, C. Chen, S. Zhou, X. Lin, W. Zuo, and L. Zhang, “Blind face restoration via deep multi-scale component dictionaries,” in *European conference on computer vision*, pp. 399–415, Springer, 2020. [8](#), [45](#)
- [29] J. Hadamard, “Sur les problèmes aux dérivées partielles et leur signification physique,” *Princeton university bulletin*, pp. 49–52, 1902. [9](#)

- [30] Z. Wang and A. C. Bovik, “A universal image quality index,” *IEEE signal processing letters*, vol. 9, no. 3, pp. 81–84, 2002. [10](#), [13](#), [55](#), [85](#)
- [31] 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. [13](#), [57](#), [85](#)
- [32] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, “Toward multimodal image-to-image translation,” *Advances in neural information processing systems*, vol. 30, 2017. [13](#), [57](#), [85](#)
- [33] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019. [13](#), [61](#), [76](#), [78](#), [85](#)
- [34] S. Baker and T. Kanade, “Hallucinating faces,” in *Proceedings Fourth IEEE international conference on automatic face and gesture recognition (Cat. No. PR00580)*, pp. 83–88, IEEE, 2000. [15](#)
- [35] S. Baker and T. Kanade, “Hallucinating faces,” in *Proceedings Fourth IEEE international conference on automatic face and gesture recognition (Cat. No. PR00580)*, pp. 83–88, IEEE, 2000. [15](#)
- [36] H. Chang, D.-Y. Yeung, and Y. Xiong, “Super-resolution through neighbor embedding,” in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1, pp. I–I, IEEE, 2004. [15](#)
- [37] X. Wang and X. Tang, “Hallucinating face by eigentransformation,” *IEEE transactions on systems, man, and cybernetics, part C (applications and reviews)*, vol. 35, pp. 425–434, 2005. [15](#)
- [38] C.-Y. Yang, S. Liu, and M.-H. Yang, “Structured face hallucination,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1099–1106, 2013. [15](#)
- [39] X. Yu and F. Porikli, “Ultra-resolving face images by discriminative generative networks,” in *European conference on computer vision*, pp. 318–333, Springer, 2016. [15](#)
- [40] J. Shi and G. Zhao, “Face hallucination via coarse-to-fine recursive kernel regression structure,” *IEEE Transactions on Multimedia*, vol. 21, pp. 2223–2236, 2019. [15](#)

- [41] Y. LeCun, L. Jackel, L. Bottou, A. Brunot, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, *et al.*, “Comparison of learning algorithms for handwritten digit recognition,” in *International conference on artificial neural networks*, vol. 60, pp. 53–60, Perth, Australia, 1995. [17](#)
- [42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, 2012. [17](#)
- [43] J. Kim, J. K. Lee, and K. M. Lee, “Deeply-recursive convolutional network for image super-resolution,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1637–1645, 2016. [17](#)
- [44] A. Esmaeilzahi, M. O. Ahmad, and M. Swamy, “Murnet: A deep recursive network for super resolution of bicubically interpolated images,” *Signal Processing: Image Communication*, vol. 94, p. 116228, 2021. [17](#)
- [45] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, “Learning face hallucination in the wild,” in *Twenty-ninth AAAI conference on artificial intelligence*, 2015. [17](#)
- [46] D. Huang and H. Liu, “Face hallucination using convolutional neural network with iterative back projection,” in *Chinese Conference on Biometric Recognition*, pp. 167–175, Springer, 2016. [17](#)
- [47] X. Chen, X. Wang, Y. Lu, W. Li, Z. Wang, and Z. Huang, “Rbpnet: An asymptotic residual back-projection network for super-resolution of very low-resolution face image,” *Neurocomputing*, vol. 376, pp. 119–127, 2020. [17](#)
- [48] T. Lu, J. Wang, J. Jiang, and Y. Zhang, “Global-local fusion network for face super-resolution,” *Neurocomputing*, vol. 387, pp. 309–320, 2020. [17](#)
- [49] K. Jiang, Z. Wang, P. Yi, T. Lu, J. Jiang, and Z. Xiong, “Dual-path deep fusion network for face image hallucination,” *IEEE Transactions on Neural Networks and Learning Systems*, 2020. [17](#)
- [50] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *2014 Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2672–2680, 2014. [18](#)
- [51] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, pp. 139–144, 2020. [18](#)

- [52] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein gan. arxiv 2017,” *arXiv preprint arXiv:1701.07875*, vol. 30, no. 4, 2017. [20](#)
- [53] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, “Least squares generative adversarial networks,” in *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802, 2017. [20](#), [21](#)
- [54] G.-J. Qi, “Loss-sensitive generative adversarial networks on lipschitz densities,” *International Journal of Computer Vision*, vol. 128, no. 5, pp. 1118–1140, 2020. [20](#)
- [55] X. Yu and F. Porikli, “Ultra-resolving face images by discriminative generative networks,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 318–333, 2016. [21](#)
- [56] S. Ko and B.-R. Dai, “Multi-laplacian gan with edge enhancement for face super resolution,” in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 3505–3512, IEEE, 2021. [22](#)
- [57] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, “Learning to super-resolve blurry face and text images,” in *Proceedings of the IEEE international conference on computer vision*, pp. 251–260, 2017. [22](#)
- [58] H. Dou, C. Chen, X. Hu, Z. Xuan, Z. Hu, and S. Peng, “Pca-srgan: Incremental orthogonal projection discrimination for face super-resolution,” in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1891–1899, 2020. [22](#)
- [59] M. Zhang and Q. Ling, “Supervised pixel-wise gan for face super-resolution,” *IEEE Transactions on Multimedia*, vol. 23, pp. 1938–1950, 2020. [22](#)
- [60] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, “Esrgan: Enhanced super-resolution generative adversarial networks,” in *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018. [22](#)
- [61] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, “Photo-realistic single image super-resolution using a generative adversarial network,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114, 2017. [22](#)
- [62] M. Zhang and Q. Ling, “Supervised pixel-wise gan for face super-resolution,” *IEEE Transactions on Multimedia*, vol. 23, pp. 1938–1950, 2020. [22](#)

- [63] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020. [23](#), [25](#), [26](#)
- [64] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep unsupervised learning using nonequilibrium thermodynamics,” in *International conference on machine learning*, pp. 2256–2265, PMLR, 2015. [23](#), [25](#)
- [65] Y. Song and S. Ermon, “Generative modeling by estimating gradients of the data distribution,” *Advances in neural information processing systems*, vol. 32, 2019. [23](#)
- [66] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley, “Face super-resolution guided by facial component heatmaps,” in *Proceedings of the European conference on computer vision (ECCV)*, pp. 217–233, 2018. [28](#), [29](#)
- [67] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, “Fsnet: End-to-end learning face super-resolution with facial priors,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2492–2501, 2018. [28](#)
- [68] D. Kim, M. Kim, G. Kwon, and D.-S. Kim, “Progressive face super-resolution via attention to facial landmark,” in *The British Machine Vision Conference (BMVC 2019)*, pp. 1–12, 2019. [28](#), [29](#)
- [69] C. Chen, X. Li, L. Yang, X. Lin, L. Zhang, and K.-Y. K. Wong, “Progressive semantic-aware style transformation for blind face restoration,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11896–11905, 2021. [28](#), [29](#)
- [70] A. Bulat and G. Tzimiropoulos, “Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 109–117, 2018. [29](#)
- [71] Y. Yu, P. Zhang, K. Zhang, W. Luo, C. Li, Y. Yuan, and G. Wang, “Multi-prior learning via neural architecture search for blind face restoration,” *arXiv preprint arXiv:2206.13962*, 2022. [29](#)
- [72] X. Hu, W. Ren, J. Yang, X. Cao, D. Wipf, B. Menze, X. Tong, and H. Zha, “Face restoration via plug-and-play 3d facial priors,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8910–8926, 2021. [29](#)
- [73] 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. [30](#), [33](#), [46](#), [47](#), [50](#), [60](#), [70](#), [79](#), [80](#), [86](#), [96](#), [98](#)
- [74] A. Brock, J. Donahue, and K. Simonyan, “Large scale gan training for high fidelity natural image synthesis,” *arXiv preprint arXiv:1809.11096*, 2018. [31](#)
- [75] A. Van Den Oord, O. Vinyals, *et al.*, “Neural discrete representation learning,” *Advances in neural information processing systems*, vol. 30, 2017. [32](#), [46](#)
- [76] 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. [33](#), [68](#), [70](#), [73](#), [85](#), [116](#)
- [77] J. Gu, Y. Shen, and B. Zhou, “Image processing using multi-code gan prior,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3012–3021, 2020. [34](#), [36](#), [37](#)
- [78] 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*, pp. 4432–4441, 2019. [36](#), [37](#)
- [79] K. C. K. Chan, X. Wang, X. Xu, J. Gu, and C. C. Loy, “Glean: Generative latent bank for large-factor image super-resolution,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14245–14254, 2021. [39](#)
- [80] X. Wang, Y. Li, H. Zhang, and Y. Shan, “Towards real-world blind face restoration with generative facial prior,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9168–9178, 2021. [39](#), [43](#)
- [81] X. Li, G. Duan, Z. Wang, J. Ren, Y. Zhang, J. Zhang, and K. Song, “Recovering extremely degraded faces by joint super-resolution and facial composite,” in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 524–530, IEEE, 2019. [45](#)
- [82] Z. Wang, J. Zhang, R. Chen, W. Wang, and P. Luo, “Restoreformer: High-quality blind face restoration from undegraded key-value pairs,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17512–17521, 2022. [46](#), [48](#)
- [83] P. Esser, R. Rombach, and B. Ommer, “Taming transformers for high-resolution image synthesis,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021. [46](#), [68](#), [71](#), [77](#), [78](#), [85](#)

- [84] X. Wang, K. Yu, C. Dong, and C. C. Loy, “Recovering realistic texture in image super-resolution by deep spatial feature transform,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 606–615, 2018. [48](#)
- [85] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Proceedings of International Conference on Computer Vision (ICCV)*, pp. 3730–3738, December 2015. [49](#)
- [86] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, “Maskgan: Towards diverse and interactive facial image manipulation,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. [50](#), [81](#), [82](#)
- [87] V. Le, J. Brandt, Z. Lin, and L. Bourdev, “Interactive facial feature localization,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 679–692, 10 2012. [50](#)
- [88] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” Tech. Rep. 07-49, University of Massachusetts, Amherst, October 2007. [51](#), [83](#)
- [89] P. Viola and M. J. Jones, “Robust real-time face detection,” *International journal of computer vision*, vol. 57, pp. 137–154, 2004. [51](#)
- [90] G. B. Huang, V. Jain, and E. Learned-Miller, “Unsupervised joint alignment of complex images,” in *ICCV*, 2007. [51](#)
- [91] G. B. Huang, M. Mattar, H. Lee, and E. Learned-Miller, “Learning to align from scratch,” in *NIPS*, 2012. [51](#)
- [92] O. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *BMVC 2015- Proceedings of the British Machine Vision Conference 2015*, British Machine Vision Association, 2015. [51](#)
- [93] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “VGGFace2: A dataset for recognising faces across pose and age,” in *International Conference on Automatic Face and Gesture Recognition*, 2018. [52](#)
- [94] K.-H. Thung and P. Raveendran, “A survey of image quality measures,” in *2009 international conference for technical postgraduates (TECHPOS)*, pp. 1–4, IEEE, 2009. [54](#)

- [95] Z. Wei, “A picture quality evaluation method based on human perception,” *Acta Electronica Sinica*, vol. 27, pp. 79–82, 1999. [54](#)
- [96] M. A. Saad, A. C. Bovik, and C. Charrier, “Blind image quality assessment: A natural scene statistics approach in the dct domain,” *IEEE transactions on Image Processing*, vol. 21, no. 8, pp. 3339–3352, 2012. [54](#)
- [97] D. M. Rouse and S. S. Hemami, “Analyzing the role of visual structure in the recognition of natural image content with multi-scale ssim,” in *Human Vision and Electronic Imaging XIII*, vol. 6806, pp. 410–423, SPIE, 2008. [55](#)
- [98] A. Hore and D. Ziou, “Image quality metrics: Psnr vs. ssim,” in *2010 20th international conference on pattern recognition*, pp. 2366–2369, IEEE, 2010. [55](#)
- [99] M. Goyal, Y. Lather, and V. Lather, “Analytical relation & comparison of psnr and ssim on babbon image and human eye perception using matlab,” *International Journal of Advanced Research in Engineering and Applied Sciences*, vol. 4, no. 5, pp. 108–119, 2015. [55](#)
- [100] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2, pp. 1398–1402, Ieee, 2003. [56](#)
- [101] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *International Conference on Learning Representations (ICLR)*, pp. 1–14, 2015. [57](#), [75](#), [77](#)
- [102] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, 2012. [57](#)
- [103] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, June 2016. [58](#)
- [104] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a “completely blind” image quality analyzer,” *IEEE Signal Processing Letters*, vol. 20, pp. 209–212, 2013. [58](#)
- [105] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, “Photo-realistic single image super-resolution using a generative adversarial network,” *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017. [59](#)

- [106] K. Nasrollahi and T. B. Moeslund, “Super-resolution: a comprehensive survey,” *Machine vision and applications*, vol. 25, pp. 1423–1468, 2014. [59](#)
- [107] K. Kwon, E. Kang, S. Lee, S.-J. Lee, H.-E. Lee, B. Yoo, and J.-J. Han, “Controllable image restoration for under-display camera in smartphones,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2073–2082, 2021. [59](#)
- [108] J. Lin, T. Zhou, and Z. Chen, “Multi-scale face restoration with sequential gating ensemble network,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018. [59](#)
- [109] A. Lahiri, S. Bairagya, S. Bera, S. Haldar, and P. K. Biswas, “Lightweight modules for efficient deep learning based image restoration,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 4, pp. 1395–1410, 2020. [59](#)
- [110] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*. Cambridge university press, 2010. [66](#)
- [111] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. [69](#), [72](#), [85](#)
- [112] J. Johnson, A. Alahi, and L. Fei-Fei, “Perceptual losses for real-time style transfer and super-resolution,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 694–711, 2016. [71](#), [75](#), [77](#)
- [113] 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. [74](#)
- [114] 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. [77](#)
- [115] S. Yang, P. Luo, C.-C. Loy, and X. Tang, “Wider face: A face detection benchmark,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5525–5533, 2016. [84](#)
- [116] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014. [85](#)

- [117] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.,” *Journal of machine learning research*, vol. 9, no. 11, 2008. [99](#)