Face Super-Resolution on Complex and Real-World Degradation Settings

by

Kyle Becker

(Under the Direction of Thirimachos Bourlai)

Abstract

Real-World Facial Super-Resolution (RWFSR) is a complex problem of producing high-resolution face images from low-resolution images captured with real-world image degradation to assist in downstream facial recognition tasks. While many approaches to RWFSR have been developed, there is a lack of hybrid models in the literature that explicitly tackle both unknown degradation estimation and identity preservation. This thesis moves toward a hybrid approach in two experiments: in the first, three state-of-the-art super-resolution algorithms are compared on benchmark training and testing datasets that are representative of simple, complex, and real-world image degradation; in the second, a novel approach is introduced that combines elements of the two best-performing algorithms from the comparison study and is evaluated on benchmark datasets as well as the MILAB-VTF(B) face image dataset. The results from these experiments reveal a trade-off between restoring images with complex degradation and maintaining identity features of face images within super-resolution models.

Index words: Artificial intelligence, computer vision, deep learning, face recognition, generative adversarial networks, real-world super-resolution

Face Super-Resolution on Complex and Real-World Degradation Settings

by

Kyle Becker

A.B. Cognitive Science, University of Georgia, 2022

A Thesis Submitted to the Graduate Faculty of the

University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Master of Science

Athens, Georgia

2024

Face Super-Resolution on Complex and Real-World Degradation Settings

by

Kyle Becker

Major Professor: Thirimachos Bourlai

Committee: Kyle Johnsen
Fred Maier

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
December 2024

Acknowledgments

I would like to express my deepest appreciation to Dr. Thirimachos Bourlai, who has not only acted as my primary advisor throughout the process of this research, but has also been invaluable to keeping me motivated and accountable so that I can complete this project. I am also thankful for the remainder of my defense committee, who have been a consistent source of support and wisdom. Special thanks to Dr. Kimberly Van Orman, whose mentorship and advice has been indispensable for me throughout my whole academic experience. I am also grateful to my peers and co-workers in the Multispectral Imagery Lab, who have been walking this path with me and keeping me sane when times got stressful. Finally, I want to thank my friends and family, especially my twin brother Drew, for reminding me what's important in life and offering their unending support throughout my graduate experience.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Real-World Face Super-Resolution (RWFSR) is a complex problem that aims to restore low-resolution (LR) face images captured in real-world scenarios into detailed high-resolution (HR) images. Super-resolution of face images is a topic of growing concern in bio-metrics and computer vision fields of research. Face recognition (FR) systems are currently being deployed in many real-world applications, including video surveillance, border control, and mobile device access (El-Naggar & Bourlai, 2019). Many FR systems function at a fixed image resolution, and those that do not typically benefit in accuracy from processing higher-resolution images, as HR images typically have more precise information than LR ones. However, real-world face images are typically highly variable and uncontrolled, with major variation in resolution, pose, occlusion, lighting, etc. Specifically in drone- or CCTV-based surveillance footage, face images are typically captured at distance and are subject to a variety of factors that degrade the quality of the image. Due to these limitations, there is an apparent need for research into image augmentation algorithms that can reconstruct sufficiently detailed face images at a resolution sufficient for modern FR systems. The reconstructed images are referred to as super-resolved (SR) images.

The domain of performing super-resolution on face images is known as face super-resolution (FSR). FSR systems are typically tailored specifically for face images, focusing on restoring a set of relevant facial features that will be recognized by FR systems. The approaches to FSR vary considerably, with some algorithms using prior embedding of facial features on which to construct SR images; and other algorithms focusing on learning the facial feature embedding directly from the training data (Jiang et al., 2021). In either case, a majority of FSR systems focus on restoring LR images with little to no degradation, typically relying on training sets of HR/LR image pairs created using simple downsampling with a bicubic kernel, or on a set of predefined image degradations. In practice, however, real-world face images are typically subject to a number of unpredictable real-world degradation parameters relating to the limitations of the camera, imaging distance, JPEG compression and other factors. Because of these limitations, many FSR techniques trained on predefined degradation settings struggle to generalize well to real-world scenarios (Philippe & Bourlai, 2024). The difference in SR performance between simple, known degradation and complex, unknown degradation is known as the domain gap.

The domain of real-world/blind super-resolution (RWSR/BSR) is focused on narrowing the domain gap and improving SR performance on images that are subject to real-world degradation. Whereas most FSR algorithms create HR/LR pairs by applying a set of known degradations to the HR images, BSR systems focus on reconstructing the SR image from the LR image where the degradation from HR to LR is unknown (Chen et al., 2022). BSR systems typically generalize well to real-world scenarios, but have one major drawback with respect to RWFSR. Most BSR systems are not specifically designed for face images, but rather for general scene and object images. This limitation means that while BSR images suffer less from the domain gap, they are not as tailored to reconstruct face-specific features that will translate well to face recognition systems.

The RWFSR task combines the challenges of face super-resolution and blind super-resolution. The goal of RWFSR is to reconstruct high quality SR images from unknown degradation scenarios that will preserve the face features and identity information of the image subject. An effective RWFSR system should be able to take LR images with unknown degradations and produce SR images that score highly on both image quality assessment metrics and face recognition accuracy. To accomplish this, it must strike the correct balance between degradation estimation and identity/feature preservation.

The content of this thesis is based around two sets of experiments:

- In the first, I collect three state-of-the-art (SOTA) image super-resolution algorithms that I believe represent the major approaches to RWFSR. I develop a methodology for fairly evaluating and comparing each algorithm's SR performance in terms of both image quality and face verification accuracy across three different degradation settings: simple, complex, and real-world. This set of experiments are referred to as the **Comparison Study**.
- In the second, I combine aspects of the best-performing super-resolution algorithms examined above to create a novel RWFSR system, with the main goal of enhancing face verification accuracy at complex and real-world degradation settings. I use the same methodology I develop above to compare my proposed algorithm against existing SOTA methods. The experiments regarding this algorithm are referred to as **IP-SCGAN**, short for *Identity Preserving, Semi Cycled Generative Adversarial Network*.

The rest of this thesis is organized as follows: the rest of this chapter, I provide background on deep-learning based computer vision, image super-resolution, and face recognition. Chapter 2 is a review of the current literature on RWFSR, including explanations for the relevant quality assessment metrics and loss functions as well as a brief discussion of the history and state-of-the-art of three different approaches to the problem. Chapter 3 discusses the overall methodology used to train, test and evaluate SR models in the experiments in this thesis, as well as a discussion around the motivation and overall goal of developing IP-SCGAN. Chapter 4 describes in detail the datasets and models used in the experiments, as well as the approaches for generating LR images of different degradation settings and the loss function framework of IP-SCGAN. Chapter 5 demonstrates the final results of quality evaluation and face recognition on the SR results from both experiments. Finally, Chapter 6 discusses the theoretical and practical implications of these results, focusing on what can be learned from these experiments and how they can be improved upon in the future.

1.1 Background

1.1.1 Computer Vision and Deep Learning

Computer vision (CV) is widely understood as the host of techniques to acquire, process, analyze, and understand complex higher-dimensional data from the the environment for scientific and technical exploration (Jähne et al., 1999). CV is typically treated as a multidimensional signal processing problem. As opposed to *time series*, the typical subject of 1-D signal processing, images typically contain data in 2 or 3 dimensions. In practice, grayscale images are typically treated as a matrix of shape $(W \times H)$, while color images are treated as matrices of shape $(W \times H \times C)$, were C represents the number of color channels and is typically 3 (i.e., RGB color space). CV tasks are typically a type of pattern recognition problem, and mirror the types of perceptual faculties that occur naturally in the human visual system, such as recovering the shape and appearance of objects in images (Szeliski, 2022). The applications of CV are incredibly varied, and include medical imaging, self-driving vehicles, surveillance, 3-D modeling, visual art, and many more (Szeliski, 2022).

While early approaches to CV tasks consisted mostly of statistical model-based approaches, recent years have seen the rise of deep learning (DL) based computer vision techniques. DL computer techniques have quickly become the state-of-the-art for many tasks, including object detection, visual tracking, semantic segmentation, and image restoration (Chai et al., 2021). DL approaches to CV are typically based on Convolutional Neural Networks (CNNs). The CNNs used in computer vision tasks such as object detection or semantic segmentation are typically derived from backbone image classification models, such as VGGNet (Simonyan & Zisserman, 2015), GoogLeNet & Inception (Szegedy, Liu, et al., 2015), and ResNet (He et al., 2016). The feature extraction capabilities of these network architectures lend themselves to be useful for deriving other CNN models for more specific tasks, such as object detection or face recognition.

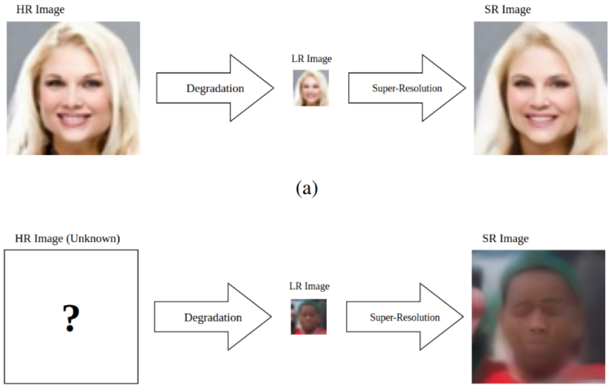
Image Synthesis

Image synthesis is a sub-set of computer vision problem that focuses on the creation of visual images, rather than detection or classification. The ability for a computer to synthesize images has many possible applications, from digital art to anti-spoofing algorithms, but perhaps the main motivation for image synthesis is to increase the amount of available training data for DL-based CV applications (Tsirikoglou et al., 2020). Image synthesis may be multi-modal, meaning other modes of information (e.g., description text) are included in the learning/synthesis pipeline, but can also operate solely in the mode of visual images.

While classical models to image synthesis exist which treat synthesis as a process of solving a potentially massive number of interdependent, high-dimensional integral equations (Tsirikoglou et al., 2020), a more popular approach in modern literature is to utilize DL to produce complete images as the output of neural networks. Two of the most popular models types for image synthesis are Variational Autoencoders (VAEs) (Kingma, 2013) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). Recent years have also seen the rise of Latent Diffusion Models (LDMs) (Rombach et al., 2022) for the synthesis of high-resolution images. While all three of these model types differ in their architectures and loss frameworks, the overall goal of all of them is for the model to learn a probabilistic data distribution such that the model can transform an input vector of random noise into a high-quality and realistic image. In addition to transforming random noise into an image, these image synthesis models are also used on input images for synthesis tasks such as restoration, super-resolution, and domain transfer. Image synthesis algorithms that take existing images as input often utilize information available within the input image, alongside statistical distributions learned from the images in the training set, to produce a transformed or enhanced version of that image. Super-resolution is one kind of task that follows this framework, as SR models use the statistical information learned from the training set as well as the distribution of information in the input LR image to generate an high-resolution SR version of the same image. The next section explores super-resolution and FSR in more detail.

1.1.2 Super-Resolution

Face super-resolution is a sub-type of image synthesis task that can be thought of as the process of recovering a corresponding HR face image from a LR face image (Jiang et al., 2021). If we let Φ be the image



(a)

(b)

Figure 1.1: The framework of the RWFSR task. (a)(top) An HR image undergoes simple bicubic downsampling before being super-resolved into the SR image. (b)(bottom) an unknown (or nonexistent) image undergoes unknown real-world degradations, and is super-resolved to estimate what the HR image would have looked like.

degradation model, we can mathematically model this problem as

$$I_{LR} = \Phi(I_{HR}, \theta), \tag{1.1}$$

where θ represents the degradation model parameters, including the blur kernel, choice of downsampling operation, and addition of noise. I_{LR} represents the LR face image, and I_{HR} represents the the original HR face image. The goal of FSR is to generate I_{SR}, or the super-resolved image, by simulating the inverse of the degradation process. If we let F represent the super-resolution model, this process can be modeled as

$$I_{SR} = F(I_{LR}, \delta) = \Phi^{-1}(I_{LR}, \delta), \tag{1.2}$$

where δ represents the parameters of the super-resolution model. In order to achieve good results on the super-resolution, δ must be optimized, which can be defined as

$$\hat{\delta} = \underset{\delta}{\operatorname{argmin}} \mathcal{L}(I_{SR}, I_{HR}), \tag{1.3}$$

where $\mathcal{L}$ represents the loss between I_{SR} and I_{HR}, and $\hat{\delta}$ is the optimal parameter set of the final model. Because $\hat{\delta}$ is calculated using $\mathcal{L}$, the choice of loss function is incredibly important to generating realistic SR images. In FSR, mean square error (MSE) loss and $\mathcal{L}_1$ loss are the most popular loss functions (Jiang et al., 2021). However, many FSR models use more sophisticated loss functions, or a combination of loss functions, as I explore in later sections.

One of the issues facing RWFSR, as well as BSR, is that in real-world scenarios, only I_{LR} is available. Therefore, to create training pairs, we must model the degradation process (Chen et al., 2022; Jiang et al., 2021). The simplest way to represent the degradation process is

$$I_{LR} = (I_{HR}) \downarrow_s, \tag{1.4}$$

where $\downarrow$ represents the downsampling operation and s represents the scale factor. To produce LR images with a simple degradation, researchers typically choose bicubic or bilinear downsampling (Hou et al., 2023; Jiang et al., 2021), though bicubic is more popular. These simple degradations poorly model real-world image degradation, which is commonly the result of complex factors such as atmospheric conditions, unfavorable lighting, and data compression (Philippe & Bourlai, 2024). Camera type and quality is also an important factor, as many real-world face images are captured on cellphone cameras, which handle image processing in ways that differ to traditional cameras (Martin & Bourlai, 2020). To mathematically capture how these complex variables impact image degradation, in many cases researchers model a more complex degradation as

$$I_{LR} = J((I_{HR} \otimes k) \downarrow_s + n), \tag{1.5}$$

where k represents the blur kernel, $\otimes$ represents a convolution operation, n represents noise, and J represents JPEG image compression (Jiang et al., 2021). By including blurring, noise, and image compression in the degradation model, LR images can be generated from corresponding HR images that more accurately model real-world degradation. Super-resolution models that handle these complex image pairs well may better generalize to real images.

1.1.3 Face Recognition

Face recognition (FR) is one of the most important applications of artificial intelligence and computer vision (X. Wang et al., 2022) that centers around using face images to identify subjects. This type of task has many potential use cases, and as such, FR has is becoming an ever-growing part of video surveillance, border control, mobile device access, and many other systems. (El-Naggar & Bourlai, 2019). Face matching in FR usually takes one of two forms (Guo & Zhang, 2019):

- Face verification (FV), in which the task is to determine wither a given pair of face images belongs to the same subject or different subjects, and
- Face identification (FI), in which the task is to match a query face to its correct identity in a set of gallery images.

While face verification is a one-to-one (1:1) matching task, face identification is a one-to-many (1:N) matching task. Each form of face verification has its own use cases; for example, a FR-based smartphone access control program might utilize face verification, whereas a criminal investigation using CCTV footage might rather use face identification, matching the query face from the footage with a gallery of face images of registered criminals.

Face recognition algorithms typically use a pipeline of related tasks to arrive at the final face matching result (Guo & Zhang, 2019). A typical FR pipeline consists of stages that handle face detection, face alignment, feature extraction (both query and gallery images), feature matching, and accuracy evaluation. While early methods used statistical methods to accomplish these steps, more recent algorithms instead use deep neural networks (DNN) to improve FR accuracy, especially in unconstrained images with large variation in pose, resolution, illumination, blur/noise, etc. (Guo & Zhang, 2019). In these deep learning-based FR methods, face matching typically looks like this:

- First, face embeddings in a deep feature space are extracted from the query face and gallery face via DNN.
- Then, a distance measure such as cosine distance is used to calculate the distance between the face embeddings of the query face and those of the gallery face.

The lower the distance between the two deep feature embeddings, the more likely the faces are to belong to the same subject.

Super-resolution can be an important pre-processing step in the FR pipeline. Since real-world FR is typically unconstrained, especially with respect to image resolution and degradation, there is a need for a reliable way to up-scale low-resolution face images such that they can be more reliably identified by FR algorithms. Since noise and synthesized artifacts from the SR process can impede the accuracy of FR algorithms, it is important for a good RWFSR system to produce visually pleasing images (i.e., minimal artifacts). However, since the real focus of FR is on extracting deep feature space embeddings from images, simply producing visually pleasing images is not enough. A successful RWFSR algorithm must also be able to generate an SR image such that the deep-feature space face embeddings can still be matched with a high-quality gallery image of the same subject. This matching process is much less about visual quality and more about how the face features are rendered in the SR image, e.g., the shape the eyes, nose and mouth, the color of the eyes, the relative position of the nose on the face, the shape of the eyebrows, etc. Thus, in order for a RWFSR algorithm to effectively assist in downstream FR tasks, it must not only create visually pleasing SR face images, but also maintain these important identity features that will lead to consistency in deep feature space face embeddings between different images of the same subject.

In the next chapter, I explore how current approaches to RWFSR have dealt with the issues of image quality and identity preservation, while also discussing the relevant metrics to determine image quality. I also discuss the loss functions that are relevant to understanding how different RWFSR approaches work.

Chapter 2

Literature Review

This chapter contains a comprehensive review of the current state of the RWFSR literature as it pertains to the experiments described in this thesis. Section 2.1 describes the foundational quality assessment metrics and network loss functions used in FSR experiments. Section 2.2 then gives an overview of the current state of research with respect to each of the approaches to super-resolution represented by the models examined in this thesis.

2.1 Metrics and Losses

This section introduces the image quality assessment (IQA) metrics I use to evaluate the SR results of the experiments, as well as a few key loss functions that are used across models in the literature on image SR. I choose a variety of universal full- and no-reference IQA metrics to quantify the overall quality of the images, as well as two face image-specific metrics to evaluate the quality of the SR face images in terms of downstream face verification.

2.1.1 Quality Assessment Metrics

PSNR

Peak signal-to-noise ratio is the most widely used full-reference image quality metric for evaluating image restoration (Chen et al., 2022). To calculate PSNR between two images, the MSE is first calculated, then the PSNR is derived,

$$\text{MSE} = \frac{1}{hwc}\|I_{SR} - I_{HR}\|_2^2, \tag{2.1}$$

$$\text{PSNR} = 10\log_{10}\left(\frac{\text{M}^2}{\text{MSE}}\right), \tag{2.2}$$

where h, w, and c represent the height, width, and channel depth of an image and M is the maximum pixel value (Jiang et al., 2021). Note that PSNR focuses on the pixelwise difference between two images; the smaller the distance, the higher the PSNR. Since PSNR focuses on the distance between every pair of pixels in the two images, it is quite good at measuring whether two images are similar; however, it does not align well with human perception, meaning it may not be a reliable metric when human perception is the main focus of the image restoration (Chen et al., 2022; Jiang et al., 2021). During experimental evaluation, I follow (Luo et al., 2020) by calculating the PSNR on the Y channel of the YCbCr color space for each image.

SSIM

Structural Similarity Index (SSIM) (Z. Wang & Bovik, 2002; Z. Wang et al., 2004) is another popular full-reference image quality metric that measures structural similarity. It does this by comparing the aspects of luminance, contrast, and structure between I_{SR} and I_{HR} as

$$\text{SSIM} = l(I_{HR}, I_{SR}) * C(I_{HR}, I_{SR}) * S(I_{HR}, I_{SR}), \tag{2.3}$$

where $l(I_{HR}, I_{SR})$ represents luminance similarity, $C(I_{HR}, I_{SR})$ represents contrast similarity, and $S(I_{HR}, I_{SR})$ represents structure similarity (Jiang et al., 2021). SSIM values vary from 0 to 1, where the higher the structural similarity of two images, the higher the SSIM value. It is reported to reflect visual quality better than PSNR (Chen et al., 2022; Z. Wang & Bovik, 2002) , making it a viable choice to pair with PSNR for image quality assessment.

LPIPS

The learned perceptual image patch similarity (LPIPS) (R. Zhang et al., 2018) is a more complex reference-based image quality metric. It uses a deep network to generate image features in deep space, then computes the l_2 distance between the two images in this deep feature space (R. Zhang et al., 2018). The more similar two images are in this deep feature space, the lower the LPIPS value. It has shown good agreement with human judgements of image quality (Chen et al., 2022; Jiang et al., 2021).

FID

Fréchet Inception Distance (FID) (Heusel et al., 2018) is a standard benchmark IQA metric for generative images. It works by using an Inception-v3 network (Szegedy, Vanhoucke, et al., 2015) to extract embeddings from I_{HR} and I_{SR} and uses Fréchet Distance (Dowson & Landau, 1982), designed to calculate the difference between two probability distributions, to find the distance between the Inception embeddings. It has a high degree of coherence with human opinion scores, meaning the lower the FID, the better the visual quality of an image.

BRISQUE

The image quality metrics discussed thus far all require SR/HR image pairs to correctly evaluate the quality of the SR images. However, in many real-world scenarios, these HR images are not available. This lack of reference images necessitates us to use a few no-reference image quality metrics to evaluate the performance of the SR algorithms on real-world images. The first of these metrics is the blind/referenceless image spacial quality evaluator (BRISQUE) (Mittal et al., 2012). BRISQUE focuses on natural scene statistics (NSS), computed in the spacial domain on locally normalized luminance coefficients. It is based on the principle that natural images contain regular statistical properties that can be measurably modified by distortions such as blurring, ringing, or noise (Shaoping Xu & Min, 2017). Using these NSS rather than distortion-specific features, BRISQUE attempts to holistically quantify the loss "naturalness" of an image due to unknown distortions. The metric was trained using a dataset of distorted images with human opinion score annotations, meaning it is "opinion-aware" (OA) (Shaoping Xu & Min, 2017) and should correlate well with human perception.

NIQE

The Natural Image Quality Evaluator (NIQE) (Mittal et al., 2013) is similar to BRISQUE (Mittal et al., 2012) in that it focuses on NSS in the spacial domain rather than information about specific distortions. However, it is "opinion-unaware" (OU) (Shaoping Xu & Min, 2017), meaning it is not trained on a database on distorted, human-rated image. In fact, it is only trained on pristine images (Mittal et al., 2013) and only makes use of measurable deviateions from statistical regularities observed in natural images (Shaoping Xu & Min, 2017). NIQE (Mittal et al., 2013) attempts to overcome the limitations inherent in "opinion-aware" models; mainly that subjective quality scores can be prone to numerous inconsistencies and that OA models generally can only assess quality degradation arising from the types of distortions present in the training images (Shaoping Xu & Min, 2017). NIQE (Mittal et al., 2013), therefore, can be thought of as "more objective" than BRISQUE (Mittal et al., 2012) in that it only considers NSS without consideration for human opinion scores.

2.1.2 Loss Functions

Adversarial Loss

Adversarial Loss was originally introduced for Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). Many SR methods employ the concept of adversarial loss, which involves two networks. In the domain of image super-resolution, the generator network G attempts to generate SR images from the given LR images, and the discriminator network D attempts to distinguish real HR images from the SR images created by the generator. A simple cross entropy-based adversarial loss can be expressed as:

$$\mathcal{L}_{\mathrm{G}}(I_{SR}) = -\log(\mathcal{D}(I_{SR})), \tag{2.4}$$

$$\mathcal{L}_{\mathrm{D}}(I_{HR}, I_{SR}) = -\log(\mathcal{D}(I_{HR})) - \log(1 - \mathcal{D}(I_{SR})), \tag{2.5}$$

where $\mathcal{L}_{\mathrm{G}}$ and $\mathcal{L}_{\mathrm{D}}$ represent the loss functions for the generator and discriminator networks, respectively, $\mathcal{D}$ represents the output of the discriminator function, and I_{HR} is randomly sampled from the HR training samples (Jiang et al., 2021).

Cycle Consistency Loss

Cycle Consistency Loss for image translation was originally introduced in CycleGAN (Zhu et al., 2020). Cycle consistency loss in the context of FSR treats super-resolution as a domain transfer problem: given the domains LR and HR, a model using cycle consistency loss tries to learn two mappings $G : LR \rightarrow HR$ and $F : HR \rightarrow LR$. Cycle consistency loss can be defined as (Jiang et al., 2021),

$$\mathcal{L}_{\text{Cycle}}(I_{LR}, I_{LR'}, I_{HR}, I_{HR'}) = \|I_{LR} - I_{LR'}\|_2 + \|I_{HR} - I_{HR'}\|_2, \tag{2.6}$$

where $I_{LR'}$ is the result of applying both mapping functions onto I_{LR}, i.e., $F(G(I_{LR}))$, and $I_{HR'}$ is the result of applying both mappings onto I_{HR}, i.e., $G(F(I_{HR})$. This loss function enforces *forward cycle consistency* (Zhu et al., 2020), whereby any image x from LR should remain the same after a full domain translation cycle, i.e., $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$. It also enforces *backward cycle consistency* (Zhu et al., 2020), whereby any any image y from HR should also remain the same after a full cycle, i.e., $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$.

Perceptual Loss

Some networks enhance the perceptual quality of an image by using a pretrained deep network to calculate the difference in perceptible information between I_{HR} and I_{SR}. Perceptual loss can be defined as

$$\mathcal{L}_{\text{Perceptual}}(I_{HR}, I_{SR}, \Psi, l) = \|\Psi^l(I_{HR}) - \Psi^l(I_{SR})\|_2, \tag{2.7}$$

where Ψ is the pretrained network and l is the l -th layer (Jiang et al., 2021). Perceptual loss essentially calculates the difference between features extracted by the pretrained model on a semantic level. This process encourages SR algorithms to create I_{SR} that is perceptually similar to I_{HR}, leading to images that typically look better to human observers but have lower PSNR values than images created with methods that leverage pixel-wise loss.

Identity Loss

Identity loss is a FSR-specific loss function that encourages SR models to preserve information that face recognition models will use to differentiate the identity of the subject. One of the most common implementations of identity loss is similar to perceptual loss and can be written as

$$\mathcal{L}_{\text{Identity}}(I_{HR}, I_{SR}, \Psi, l) = D\big(\Psi^l(I_{HR}), \Psi^l(I_{SR})\big), \tag{2.8}$$

where Ψ is a pretrained face recognition network and l is the l -th layer. Here, D represents a distance metric between the FR embeddings of I_{HR} and I_{SR}. The distance metric between the embeddings may differ depending on the implementation or FR network used, but cosine distance is among the most common (Dastmalchi & Aghaeinia, 2022; Deng et al., 2019; Meng et al., 2021). Identity loss encourages the network to produce SR images with similar face recognition embeddings to their HR counterparts, essentially preserving the identity information within a face image.

ArcFace Loss

ArcFace loss (Deng et al., 2019) is a loss function used for face representation learning and face recognition (Nalty et al., 2022). It is one of the most widely adopted loss functions for this task, as it has displayed SOTA performance on a number of popular benchmarks (Meng et al., 2021). It can be defined as follows (Meng et al., 2021): suppose a training batch of N face samples $\{f_i, y_i\}_{i=1}^{N}$ of n unique identities, where $f_i \in \mathbb{R}^d$ represents the d-dimensional embedding output from the last fully connected layer of the network and $y_i = 1, \ldots, n$ is its associated class label. ArcFace (Deng et al., 2019) makes this face feature embedding more discriminative between classes by optimizing it on a hypersphere manifold. Using this hypersphere, ArcFace (Deng et al., 2019) can define the angle θ_j between f_i and the j-th class center $w_j \in \mathbb{R}^d$ as $w_j^T f_i = \|w_j\| \|f_i\| \cos\theta_j$. With this in mind, Arcface Loss can be formulated as

$$\mathcal{L}_{\text{ArcFace}} = -\frac{1}{N} \sum_{i=1}^{n} \log \frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)} + \sum_{j \neq y_i} e^{s\cos\theta_j}}, \tag{2.9}$$

where $m > 0$ is the additive angular margin and s is the scaling parameter. **MagFace** (Meng et al., 2021) extends this loss function further by encoding a quality measure into the face representation by optimizing over the magnitude $a_i = \|f_i\|$ without feature normalization. MagFace (Meng et al., 2021) can thus introduce a magnitude-aware angular margin $m(a_i)$ and regularizer term $g(a_i)$ that push higher quality samples toward the center of their respective class clusters while pushing away harder faces (Nalty et al., 2022). The feature magnitude calculated by MagFace is reported to be a good metric for face quality as well (Meng et al., 2021), especially when paired images are not available for face verification. By replacing the additive angular margin m in (2.9) with $m(a_i)$ and adding the regularizer term, MagFace (Meng et al., 2021) loss can be defined as

$$\mathcal{L}_{\text{MagFace}} = -\frac{1}{N}\sum_{i=1}^{n} \log \frac{e^{s\cos(\theta_{y_i}+m(a_i))}}{e^{s\cos(\theta_{y_i}+m(a_i))} + \sum_{j\neq y_i} e^{s\cos\theta_j}} + \lambda_g g(a_i), \tag{2.10}$$

where λ_g is a hyper-parameter used to trade-off between classification and regularization losses.

2.2 Approaches to Super-Resolution

The models used in this paper span across three major domains of SR algorithms: **degradation modeling-based methods** (Bell-Kligler et al., 2019; Chen et al., 2022; Gu et al., 2019; Luo et al., 2020, 2021), **domain translation-based methods** (Bulat et al., 2018; Chen et al., 2022; Hou et al., 2023; Y. Zhang et al., 2020; Zhu et al., 2020), and **identity-preserving methods** (Dastmalchi & Aghaeinia, 2022; Hsu et al., 2019; Huang et al., 2017, 2019; Jiang et al., 2021). In this section, I define each of these domains and offer an overview of related research in each domain.

2.2.1 Degradation Modeling Methods

One of the main problems facing real-world SR is that in many real cases, the degradation model (i.e., Φ from Eq. (1.1)) is unknown (Chen et al., 2022). In order to deal with this problem, many solutions attempt to approximate the degradation model and parameters. These solutions are referred to as degradation modeling methods. These methods aim to reverse unknown image degradation by modeling it using an equation such as Eq. (1.5) and approximating the blur kernel, downsampling operation, and noise input. Once these parameters are approximated, they are inversely applied to the LR image to create the SR image. Some methods discussed also use intermediate SR results to refine the degradation model approximation, leading to better SR results with each iteration of the algorithm. The degradation modeling methods discussed here largely do not explicitly try to model image compression.

Early degradation modeling-based methods model image SR as an energy minimization problem. These types of methods utilize numerical optimization to perform image SR, and alternately optimize the blur kernel and intermediate SR image for a certain number of iterations (Chen et al., 2022; Shao & Elad, 2015). The blur kernel estimation function can be expressed mathematically as

$$\min_{I_{SR},k} \lambda \|SKI_{SR} - I_{LR}\|_2^2 + \mathfrak{R}(I_{SR}, k) + \eta \|KI_{SR} - \tilde{I}_{SR}\|_2^2 \tag{2.11}$$

where λ and η are term-balancing parameters. I_{SR} here denotes the intermediate super-resolved image, and $\tilde{I}_{SR}$ represents the super-resolved image generated by a non-blind learning-based method. K here represents the blur matrix corresponding to the blur kernel k. $\mathfrak{R}(I_{SR}, k)$ represents the direct bi-l_0-l_2-norm regularization term for the intermediate SR image and blur kernel, which aids in accurate estimation of the blur kernel (Shao & Elad, 2015). The main contribution of this work is that it demonstrates that iterative optimization of the SR image and blur kernel can lead to pleasing results in complex or real-world degradation scenarios.

Further research has extended image SR past numeric optimization, using deep neural networks to learn a representation of the degradation model and parameters (Bell-Kligler et al., 2019; Gu et al., 2019; Luo et al., 2020, 2021). KernelGAN (Bell-Kligler et al., 2019) introduces an internal generative adversarial network (GAN) that learns a unique internal distribution of patches from a single LR training image, leveraging the assumption that the best kernel for SR will be the one that best preserves the patch distribution across different scales of the LR image. The objective function of KernelGAN (Bell-Kligler et al., 2019) can be defined as:

$$G^*(I_{LR}) = \underset{G}{\operatorname{argmin}} \max_{D} \left\{ \mathbb{E}_{x \sim \text{patches}(I_{LR})}[|D(x) - 1| + |D(G(x))|] + \mathcal{R} \right\} \tag{2.12}$$

where G and D represent the generator and discriminator, respectively; and $\mathcal{R}$ is the regularization term on the downscaling kernel resulting from G. Here, G^* represents the ideal SR downscaling function for a given LR image, which is the result of the generator network converging. While KernelGAN (Bell-Kligler et al., 2019) is reported to perform well on many benchmark image SR datasets, some authors (Gu et al., 2019) note that internal learning-based methods are easily susceptible to noise in the input image. These authors instead propose a method called Iterative Kernel Correction (IKC) (Gu et al., 2019). The IKC algorithm is founded on the observation that kernel mismatch between the SR model and actual image can result in artifacts such as over-sharpening or over-smoothing. These artifacts can in turn be used to iteratively correct the blur kernel estimation. In a nutshell, IKC uses three models: a SR model $\mathcal{F}$, a predictor $\mathcal{P}$, and a corrector $\mathcal{C}$. Dimensionality reduction is performed on the blur kernel, and the result is represented by h. The method begins by producing an initial estimation h_0 from the predictor function $h_0 = \mathcal{P}(I_{LR})$. This initial estimation is then used to produce the first SR result $I_0^{SR} = \mathcal{F}(I_{LR}, h_0)$. The algorithm then uses these initial estimations to iteratively correct the estimated blur kernel. At the i-th iteration, the correction process can be modeled as:

$$\Delta h_i = \mathcal{C}(I_i^{SR}, h_{i-1}) \tag{2.13}$$

$$h_i = h_{i-1} + \Delta h_i \tag{2.14}$$

$$I_i^{SR} = \mathcal{F}(I_{LR}, h_i), \tag{2.15}$$

where h_{i-1}, Δh_i, and h_i denote the previous estimation, correcting update, and new estimation respectively; I_i^{SR} denotes the new SR result at the current iteration.

While IKC offers good performance on many benchmark image SR datasets (Bell-Kligler et al., 2019), some authors (Luo et al., 2020, 2021) note that it requires the training of multiple disconnected networks. This framework creates a limitation, since in practice these networks may be incompatible with each other, causing errors to compound to produce an inaccurate kernel estimation. To address this issue, they introduce the Deep Alternating Network (DAN) (Luo et al., 2020). Similar to IKC, DAN (Luo et al., 2020) breaks down image SR into two steps: estimating the blur kernel k and applying it to generate the SR image I_{SR}. However, DAN (Luo et al., 2020) alternately optimizes the blur kernel and restores the SR image with a process that differs from IKC. At each iteration i this process can be modeled as

$$\begin{cases} k_{i+1} = \underset{k}{\operatorname{argmin}} \, \| I_{LR} - (I_i^{SR} \otimes k) \downarrow_s \|_1 \\ I_{i+1}^{SR} = \underset{I_{SR}}{\operatorname{argmin}} \, \| I_{LR} - (I_{SR} \otimes k_i) \downarrow_s \|_1 + \phi(I_{HR}) \end{cases}, \tag{2.16}$$

where $\phi(I_{HR})$ is the prior term for the HR image, which is usually unknown and has no analytic expression. DAN (Luo et al., 2020) uses two deep networks to solve each half of this equation, known as the *Estimator* and *Restorer*. While IKC similarly uses three disconnected models, not all of which take the

LR image as input, DAN joins its two models together to form an end-to-end trainable network. This end-to-end framework allows both Estimator and Restorer to learn information from the input image. In addition, the Restorer is trained with the kernel estimated by the Estimator; since the two networks are connected and learn together, the Restorer is thus more tolerant to estimation error in the blur kernel during testing.

2.2.2 Domain Translation Methods

While degradation modeling methods demonstrate SOTA performance on blind SR tasks (Bell-Kligler et al., 2019; Gu et al., 2019; Luo et al., 2020, 2021), they suffer from one major limitation. Yuan et al., 2018 notes that while many blind SR algorithms rely on paired HR/LR images to model real-world image degradation, these types of image pairs largely do not exist for real world images. To address this issue, these authors decide to explore an unsupervised learning approach, drawing heavily from CycleGAN (Zhu et al., 2020) and its effectiveness at performing image-to-image domain translation when trained on unpaired data. Image SR, however, is different a different problem than domain translation in two key ways (Yuan et al., 2018):

- First, domain translation focuses on converting an image from one domain into an image of the same size from another domain, whereas SR requires the generated image to be larger (higher resolution) than the input image.

- Second, image SR requires the generated image to be of higher quality (i.e. less noise, blur, etc.) than the input image. Domain translation generally does not account for image quality; in face,

> Yuan et al., 2018 notes that directly upsampling and translating an LR image to construct an SR one would lead to an enlargement of noisy patterns, making the image low quality.

These authors take multiple steps to alleviate these issues. First, they formulate the SR problem as a more general version than defined in Eq. (1.1), which can be written as

$$I_{LR} = f_n(f_d(I_{HR})) + n, \tag{2.17}$$

where f_d is an unknown downsampling process, and f_n is a degradation function that may introduce complex noises, shift, blur, and compression. Next, they introduce their framework of Cycle-in-Cycle GAN (CinCGAN) based on two coupled CycleGANs in an end-to-end framework. CinCGAN proposes three image domains X, Y and Z, representing LR images, "clean" LR images created from directly downsampling HR images without noise, and HR images, respectively. The goal of CinCGAN is to learn a mapping from X to Y and Y to Z, i.e., from LR to "clean" LR, then from "clean" LR to HR. The adoption of 3 domains allows the super-resolution problem to be tackled in domain translation steps, where the $X \rightarrow Y$ translation represents image de-noising and the $Y \rightarrow Z$ translation represents upsampling. CinCGAN is trained with a sample of unpaired images $x_i \in X, y_j \in Y$, and $z_j \in Z$, where y_j is downsampled from z_j using a bicubic kernel (Yuan et al., 2018). A key feature of this framework is that while images in Y are synthesized from images in Z, there is no pair-wise relationship between the domains X and Z, i.e., the real-world LR and HR image domains. As such, the CinCGAN framework is able to train on unpaired HR and LR image datasets. In later research (R. Zhang et al., 2018), these authors apply CinCGANs to improve SR performance and introduce different upscaling factors.

While the methods discussed to this point have been SR methods for general scene images, there is also prior research focusing on face images. LRGAN (Bulat et al., 2018) is one of the first models to apply a cycle-consistency framework to face images, using two sub-networks to uphold forward and backward cycle consistency: a "learning to degrade" sub-network to learn the translation from HR to LR (backward cycle consistency) and a "learning-to-SR" sub-network to learn the translation from LR to SR (forward cycle consistency). These advances inspired Hou et al., 2023 to construct a model using a similar two-sub-network structure in a framework they call Semi-Cycled GAN (SCGAN). SCGAN differs from the LRGAN framework in that its two sub-networks are coupled, meaning they each consist of two branches, but one branch is shared between the two networks. The shared branch is an image restoration branch, while the other two are a synthetic HR image degradation branch and a real-world HR face degradation branch. The synthetic HR image degradation branch is coupled with the image restoration branch to enforce forward cycle consistency, while the real-world HR image degradation branch is coupled with the restoration branch to enforce backward cycle consistency. It has shown superior performance on FSR tasks to that of both CycleGAN and LRGAN on benchmark datasets (Hou et al., 2023). The architecture of this model is discussed in further detail in section 4.3.

2.2.3 Identity-Preserving Methods

One key difference between face SR and general image SR is the fact that face images contain identity information, i.e., features of the image that identify the face as belonging to one individual. An ideal FSR system is not just one that produces visually pleasing results, but also one that accurately maintains or reconstructs a subject's identity features for use in downstream face recognition tasks (Jiang et al., 2021). There are typically two ways of accomplishing this: some methods (Hsu et al., 2019) use a pairwise-based data framework, while others (Dastmalchi & Aghaeinia, 2022; Huang et al., 2017, 2019) explicitly incorporate FR models into the SR framework. Both approaches are discussed here.

Siamese GAN (SiGAN) (Hsu et al., 2019) represents one of the first works in targeting identity preservation in face SR. Develop a GAN framework composed of two identical generators and one discriminator. They use a pairwise data scheme for network training, meaning the training images are presented to the network in pairs. Each training pair comes with a binary label indicating whether the face pair are the same or different identities. In addition to adversarial and reconstruction loss to learn the super-resolution process, SiGAN also factors in a contrastive loss term between the pairs of generated samples that help it to learn differentiating identity features of the face images.

Pairwise data can be difficult to find or time-intensive to produce. Instead, some authors (Dastmalchi & Aghaeinia, 2022; Huang et al., 2019) use a framework to directly enforce identity preservation without relying on pairwise data. These frameworks use wavelet transform (WT) to preserve textural and contextual details while explicitly incorporating identity-based loss functions to preserve important identity features of the image. Wavelet transform is a popular tool for image SR, as it has been shown to be an effective at representing multi-resolution images (Akansu & Haddad, 2001) and can depict textural and contextual information in an image at different scales and resolutions (Huang et al., 2017). Wavelet-SRNet (WaSRNet) (Huang et al., 2017) represents one of the first algorithms to utilize WT in the image SR task. These authors frame the image SR problem as a wavelet-coefficient prediction task and use a CNN to learn to predict the wavelet coefficients of the SR image from the LR image, maintaining the textural features and global topology of face images during SR. Wavelet-SRGAN (WaSRGAN) (Huang et al., 2019) extends this concept further by treating the wavelet-prediction CNN from WaSRNET as a generator network in a GAN framework and introduces a wavelet-domain discriminator network to improve the perceptual quality of the generated SR images. WaSRGAN also introduces explicit identity verification into the WT SR framework to directly enforce identity preservation. The most recent addition to this lineage of SR algorithms, the Wavelet Integrated, Identity Preserving Adversarial Network (WIPA) Dastmalchi and Aghaeinia, 2022, further refines the use of both WT and face identity loss. WIPA incorporates Wavelet Prediction blocks into a baseline deep network, resulting in a network architecture that learns both image features and wavelet-domain feature maps. To increase perceptual quality and identity preservation, WIPA also includes a discriminator network and identity loss using SphereFace W. Liu et al., 2017 to calculate the cosine distance between the deep face features of the generated SR image and HR training image. WIPA has demonstrated SOTA results on identity-preserving face SR of synthetic LR images produces by bicubic downsampling (Dastmalchi & Aghaeinia, 2022); however, it fails to explicitly account for different degradation scenarios, meaning these results may not generalize to complex or real-world degradation.

Chapter 3

Methodology

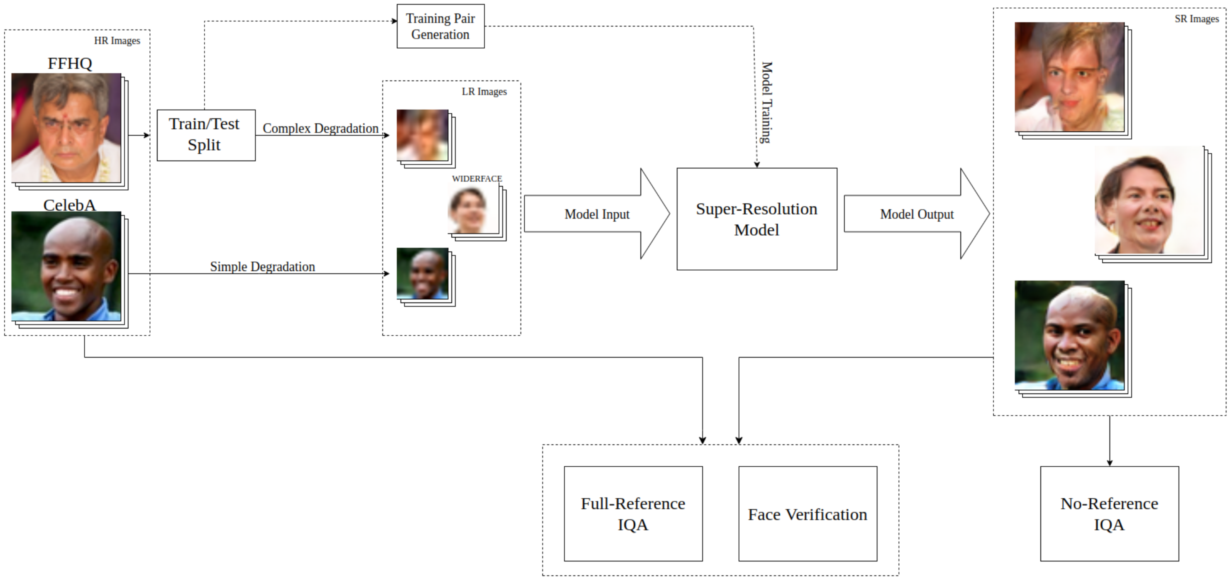


Figure 3.1: The methodology for the model evaluation described in the studies in this thesis.

This chapter describes the overall methodology and framework behind the experiments described in the next chapter. To begin, I provide an overview of the experimental methodology for evaluating the performance of each SR model on simple, complex, and real-world image degradation. I then describe how the testing images for each degradation setting are generated.

3.1 Overview

For ensure valid results for both the comparison study and the evaluation of IP-SCGAN, it is imperative to design an experimental methodology that is consistent, thorough, and repeatable. The overall methodology outlined here for model evaluation details the overall framework for training and evaluating each model. In the comparison study, the methodology is applied to each model, and the results are evaluated against each other. In the evaluation of IP-SCGAN, the methodology is applied to the novel method, and its results are compared against the SOTA models used in the comparison study. A graphical version of the methodology is pictured in Figure 3.1.

The methodology used here is derived from Hou et al., 2023. To begin, our selected datasets are train/test split to comprise our training and testing datasets. The training set is set aside to be used for model training, while the testing sets undergo one of 2 synthetic degradation settings, outlined in Section 4.2. A third testing set of real-world LR images is also included in the testing set, which does not require any synthetic degradation. Each model is trained on the training set of images using its prescribed training procedure. After training, images from the simple, complex, and real-world degradation settings are super-resolved using each model.

The SR results for each dataset are evaluated based on two sets of IQA metrics. Full-reference (FR) statistics evaluate the quality of each SR image in relation to its corresponding HR image. FR statistics generally measure the difference between a degraded image and its pristine counterpart. No-reference (NR) IQA statistics evaluate the quality of each SR image without a pristine HR counterpart, making them extremely useful in cases where no HR image is available, such as in real-world settings. In these experiments, the SR results for the simple and complex degradation settings are evaluated based on FR

IQA statistics, and identity preservation is evaluated by performing 1:1 face verification using ArcFace (Deng et al., 2019). For all images, the SR results are also evaluated using a set of NR IQA statistics, using MagFace score (Meng et al., 2021) as a proxy for identity preservation. The conclusions I draw about model performance are based around weighting the scores of each full- and no-reference IQA statistic, face verification scores, and subjective opinion scores about the quality of the images produced.

3.2 Developing IP-SCGAN

The methods investigated in the comparison study each have strengths and weaknesses with regards to the RWFSR task. Motivated by a lack of RWFSR methods in the literature that attempt to explicitly incorporate aspects of BSR and FSR, I decided to create a hybrid method of my own by combining aspects of SCGAN (Hou et al., 2023) and WIPA (Dastmalchi & Aghaeinia, 2022). Each of these algorithms have strengths and tradeoffs that make them good candidates for combining into a hybrid method:

- SCGAN (Hou et al., 2023) specializes in generating high perceptual quality face images from LR images with complex or real-world degradation, but performs poorly at identity preservation during the SR process.

- WIPA (Dastmalchi & Aghaeinia, 2022) performs extremely well at identity preservation on LR images with simple degradation, but overall model performance drops significantly at complex and real-world degradation settings.

I propose a hybrid model, combining elements of both of these state-of-the-art solutions, which I call Identity Preserving, Semi-Cycled Adversarial Network (IP-SCGAN). IP-SCGAN retains most of the same framework from SCGAN (Hou et al., 2023), but leverages a SpereFace network (W. Liu et al.,

2017) to incorporate an explicit identity loss term during model training, similar to WIPA (Dastmalchi & Aghaeinia, 2022). The specifics of the IP-SCGAN framework and architecture are discussed in the next chapter. Here, I discuss the overall structure and goals of the IP-SCGAN study.

To evaluate the effectiveness of IP-SCGAN, I use the same methodology outlined in 3.1. My IP-SCGAN experiments mainly consist of three parts:

- I begin by conducting an ablation study in order to determine the optimal weighting of the new identity loss term, comparing results directly with the baseline SCGAN.
- After determining the optimal weighting for the identity loss term, I retrain and evaluate the IP-SCGAN model on the same training and testing datasets from the comparison study.
- As a final test, I use face recognition to evaluate IP-SCGAN, SCGAN, and WIPA on MILAB-VTF(B), a real-world, identified dataset of face images taken outdoors from various distances.

The specifics of both of these experiments are discussed further in the next chapter. When evaluating model results for the IP-SCGAN study, I place more emphasis on face verification results than on overall image quality metrics. Since the SCGAN framework has already been demonstrated to perform quite well at creating high perceptual quality images during FSR (Hou et al., 2023), the main goal of the IP-SCGAN model is to increase face verification accuracy at complex and real-world degradation settings while maintaining competitive performance in terms of IQA metrics.

Chapter 4

Experiments

This chapter describes the specific implementation of the experimental methodology I outlined in the last chapter. I begin by describing the datasets I use for model training and evaluation. Next, I discuss in detail the architectures of each SR model I chose to compare in the comparison study, before describing how I took aspects of the different models and combined them into IP-SCGAN, a novel face SR framework that incorporates both domain translation and identity preservation aspects into the SR problem.

4.1 Datasets

4.1.1 CelebA

CelebFaces Attributes Dataset (CelebA) (Z. Liu et al., 2015) is a publicly available, large-scale face attributes dataset consisting of 202,599 face images belonging to 10,177 identities. Each face image includes annotations of 5 landmark locations and 40 binary attributes, which largely are not used in this study. Images in this dataset contain large variations in terms of pose and background clutter. Following Hou et al., 2023, I randomly sample 5,000 faces from CelebA to use as a testing set. To align the CelebA images consistently with the FFHQ alignment, I use face alignment scripts provided in Back, 2021, published publicly on Github.

4.1.2 FFHQ

Flickr-Faces HQ (FFHQ) is a large, publicly-available, high-quality face image dataset originally created as a benchmark for GANS in Karras et al., 2019. It contains roughly 70K PNG images at 1024×1024 resolution and contains large variation in age, ethnicity, and image background, as well as accessories such as eyeglasses, hats, etc. All images were crawled from Flickr and automatically aligned and cropped using the Dlib C++ library. Following Hou et al., 2023, I randomly sample 20,000 face images to be used as the training set and 5,000 additional face images to be used as a testing set.

4.1.3 WIDER FACE

WIDER FACE (Yang et al., 2016) is a face detection benchmark dataset consisting of 32,203 images with a total of 393,703 faces with large variation in scale, pose and occlusion. Images were selected from the publicly-available WIDER dataset (Xiong et al., 2015). I use the same sample of 2,000 low-resolution faces as used in Hou et al., 2023 as a real-world testing set.

4.1.4 MILAB-VTF(B)

The MILAB-VTF(B) dataset (Bourlai et al., 2024) is a multi-distance, unconstrained thermal-visible face image dataset. It consists of 400 subjects, imaged using both visual- and thermal-band cameras. Video frames of each subject are captured both indoors at a distance of 1.5m and outdoors at distances of 100, 200, 300, and 400m and feature variable pose, illumination, location , expression, and occlusion. I utilize this dataset to get representative results for an actual real-world face recognition problem for my evaluation of IP-SCGAN. Following Philippe and Bourlai, 2024, I collect my test set from 80 subjects reserved for testing purposes, focusing on the 300m outdoor, visible setting. I use the FFHQ face alignment script provided by Back, 2021 to align, crop, and resize face images to a resolution of 16×16 for FSR comparison. Due to the alignment script's difficulties at detecting faces at such real-world degradation, a large portion of the faces in the dataset were not detected and therefore pruned out; I was left with roughly 230 images corresponding to 46 identities in the LR setting. For the HR images, I take the indoor frames of the 46 subjects collected indoors at 1.5m. Using the same face alignment script, I align, crop, and resize the face images directly to a resolution of 112×112 to be compatible for ArcFace (Deng et al., 2019) face recognition.

4.2 Image Degradation

This section describes the image degradation procedure that is used to obtain each testing set of images. Testing images are categorized in one of three ways: simple, complex, and real-world degradation. For each degradation setting, only a single source dataset is used. The specific datasets are described in the next chapter.

4.2.1 Simple Degradation

The simple degradation testing set used in the experiments in this thesis were taken from CelebA (Z. Liu et al., 2015). After aligning the images with FFHQ and resizing to a 64×64 resolution, I create the testing

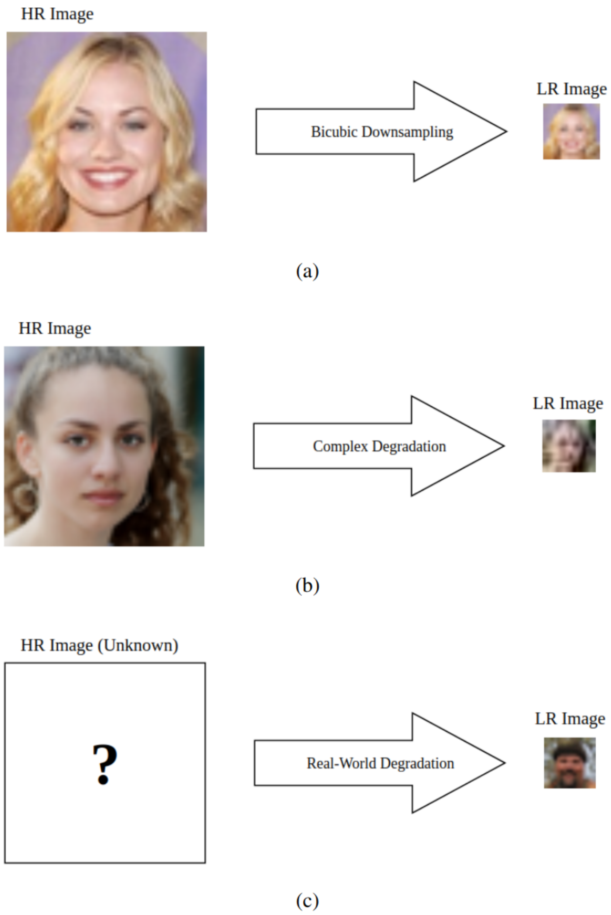


Figure 4.1: An illustration of face images undergoing different degradation settings. (a)(top) An image from CelebA, before and after bicubic downsampling. (b)(middle) An image from FFHQ, before and after our complex degradation algorithm. (c)(bottom) An image from WIDER FACE after real-world degradation, with indication that the ground-truth HR image is unknown.

set by further downsampling the images with a scale factor of 4 and a bicubic kernel, resulting in a testing set of 16×16 resolution images with simple degradation.

4.2.2 Complex Degradation

For the complex degradation testing set, I start with a set of 5,000 HR faces sampled from FFHQ (Karras et al., 2019). After resizing the HR images to a 64×64 resolution I follow Hou et al., 2023 by applying the following degradation formula to each image:

$$I_{LR} = ((I_{HR} \otimes k_r) \downarrow + n_\sigma)_{JPEG_q}, \tag{4.1}$$

where where k_r represents a Gaussian blur kernel with radius r, $\downarrow$ denotes downsampling by a factor of 4 with random choice between bicubic and bilinear kernel, n_σ represents additive Gaussian noise with a standard deviation σ, and $JPEG_q$ denotes JPEG compression with a quality factor q. For each image in the testing set, I randomly sample $r \in [0.2, 4.0]$, $\sigma \in [1, 25]$, and $q \in [30, 95]$. In general, the image degradation achieved by sampling these values was quite aggressive, and in many cases all models performed worse on complex degradation than on real-world degradation. These findings suggest that the values chosen may not adequately represent a "middle ground" between simple and real-world degradation; however, the images produced from this degradation process were still useful in that I could use them to evaluate model performance on aggressive image degradation while still allowing for ground-truth comparison and full-reference IQA statistics.

4.2.3 Real-World Degradation

The real-world degradation testing set is made up of 5,000 images from the WIDER FACE (Yang et al., 2016) dataset. These images were captured "in the wild" and are already the result of unknown, real-world degradation processes. While these images make for the best representation of real-world degradation available in these studies, the major drawback of WIDER FACE is that it does not have ground-truth or high-quality images of the faces available. Due to this limitation, I was only able to evaluate model performance on *true* real-world degradation using no-reference statistics. This is particularly troublesome when trying to evaluate identity preservation, as there were no high-quality face images to allow me to calculate face verification accuracy. However, no-reference IQA statistics should still provide a way to compare model performance on image quality for this dataset, and MagFace (Meng et al., 2021) scores offer a no-reference method for comparing identity preservation across different models.

4.3 Models

In this section, I describe the architecture of our chosen models, as well as the general training procedure for each model.

4.3.1 DAN

In our experiments, I use DANv1, proposed in Luo et al., 2020[1]. DAN is composed of two networks, *Estimator* and *Restorer*. These networks are made up of conditional residual blocks (CRB) (Luo et al., 2020) that are based on residual blocks for image SR proposed in Y. Zhang et al., 2018, which can be modeled as

$$f_{out} = R(Concat([f_{basic}, f_{cond}])) + f_{basic}, \tag{4.2}$$

where f_{out} is the output of the block, f_{basic} is the basic input, and f_{cond} is the conditional input. $Concat$ represents concatenation, and R denotes the residual mapping function, which is composed of two 3×3 convolutions layers plus a channel attention layer (Hu et al., 2018). This architecture is shown in Figure 4.3.

[1] At the time of conducting these experiments, these authors have released a new version of this model, DANv2 (Luo et al., 2021). However, due to bugs in the officially released implementation of this model, DANv2 was incapable of being correctly trained; for this reason, DANv1 is selected over DANv2.

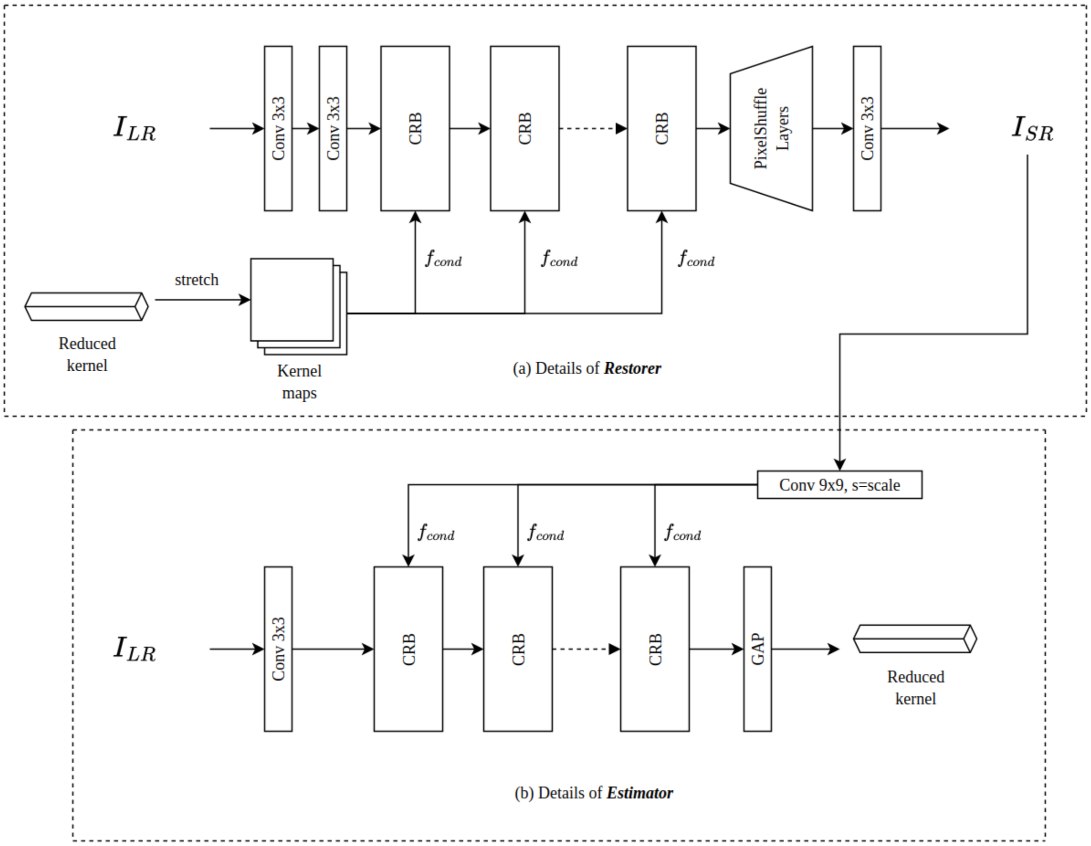


Figure 4.2: Architectures of *Restorer* and *Estimator* used by DANv1.

The *Estimator* network that estimates the degradation kernel begins by downsampling an SR image by a convolution layer with a stride s, typically corresponding to the scale of the SR. These feature maps serve as the conditional inputs for all the CRBs in *Estimator*, while the LR image serves as the basic input to the network. The network ends with a global average pooling layer to squeeze the features and form the predicted kernel. The network consists of a total of 5 CRBs, all with 32-channel basic and conditional inputs.

The *Restorer* takes the kernel estimated by *Estimator* and stretches it to the same spatial dimension as the LR image. This stretched kernel is used as the conditional input to all the CRB layers. The network

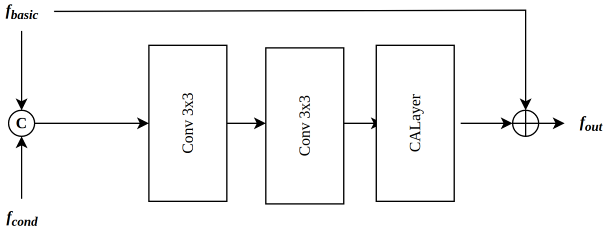
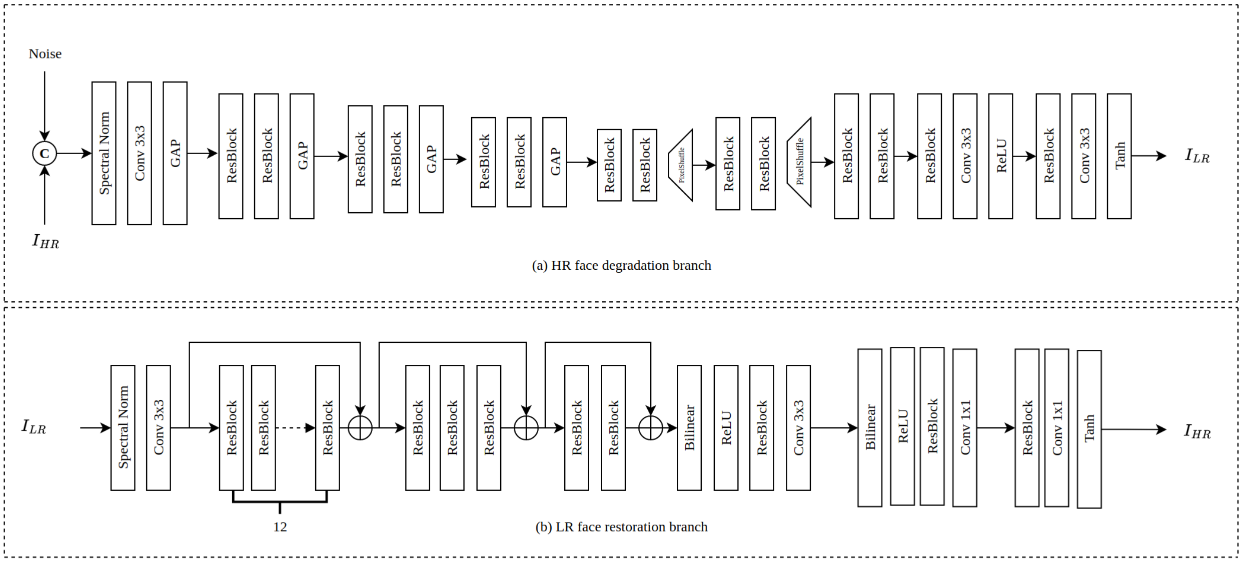


Figure 4.3: Architecture of the CRB block used by DANv1.

ends as with PixelShuffle (Shi et al., 2016) layers to upscale the features to the final SR size. In total, *Restorer* has 60 CRBs with 64-channel and basic inputs and 10-channel conditional inputs. Note that the 64-channel basic input is seemingly too large to suit 16×16 images, but the results of applying this architecture on smaller face images was still competitive with other methods The architecture of both *Estimator* and *Restorer* are shown in Figure 4.2. Since *Restorer* creates the conditional input for *Estimator* and vice versa, a full model can be built by successively alternating *Estimator* and *Restorer* networks to create a full, end-to-end trainable network for SR.

4.3.2 SCGAN

The framework of SCGAN (Hou et al., 2023) is designed specifically to enforce forward and backward cycle consistency during training on unpaired images. It consists of two coupled networks made up of 2 branches each; in total, there are 3 branches. The branch shared between the two networks is an LR restoration branch ($\mathcal{R}_{LS}$), and each network additionally has an independent HR degradation branch with an encoder-decoder architecture. One independent HR degradation branch learns how to degrade the HR images from the training dataset into the LR domain ($\mathcal{D}_{HL}$), and the other learns how to degrade the generated SR images back into the real-world LR images from the training set ($\mathcal{D}_{SL}$). The LR restora-



(a) HR face degradation branch

(b) LR face restoration branch

Figure 4.4: Architectures of the network branches used in SCGAN.

tion branch is shared between the two networks to learn the ultimate translation from LR to HR in order to perform the super-resolution.

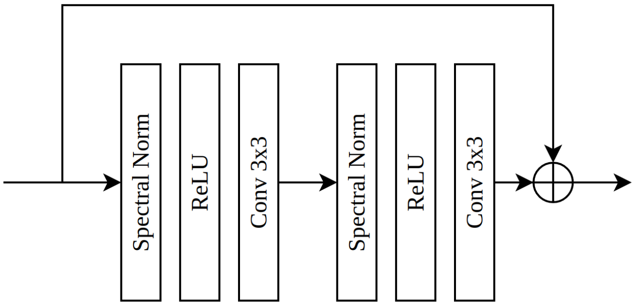


Figure 4.5: Architecture of the ResBlocks used in SCGAN.

Both the HR face degradation and LR face restoration branches make liberal use of residual blocks (ResBlocks), described by Hou et al., 2023. The ResBlock architecture is shown in Figure 4.5. The HR face degradation branches share the same encoder-decoder architecture, which is shown in Figure 4.4.

The encoder starts with a Spectral Normalization (SN) layer—originally proposed by Miyato et al., 2018—along with 3×3 convolution layer and a global average pooling (GAP) layer. The rest of the encoder architecture consists of 6 ResBlocks with a GAP layer after every 2 ResBlocks to downsample the image or encoded feature map. The decoder has a similar architecture, beginning with 6 ResBlocks, but has a PixelShuffle (Shi et al., 2016) layer after the second and fourth ResBlocks to upsample the image or encoded feature map. After the sixth ResBlock, the decoder finishes with two groups, each consisting of a ResBlock, a 3×3 convolution layer, and either ReLU or Tanh activation to output the final LR image. The LR face restoration branch similarly begins with an SN layer and 3×3 convolution layer, but then goes to 3 groups of ResBlocks, with 12, 3, and 2 ResBlocks respectively in each group. Each group of ResBlocks is also skipped by a skip connection to preserve high-frequency features. The architecture finishes with two blocks composed of a binlinear upsampling layer, ReLU activation, ResBlocks and convolution layers before concluding with a Tanh activation layer to output the final SR image.

While training, the model takes as input two unpaired images: the real-world LR image $\mathbf{I}_{rL}$ and the HR image $\mathbf{I}_{rH}$ and produces 4 result images. From $\mathbf{I}_{rH}$, $\mathcal{D}_{HL}$ and $\mathcal{R}_{LS}$ work together to produce a synthesized LR image $\mathbf{I}_{sL}$ and a synthesized SR image $\mathbf{I}_{sS}$. From $\mathbf{I}_{rL}$, $\mathcal{D}_{SL}$ and $\mathcal{R}_{LS}$ work together to produce a real-world SR image $\mathbf{I}_{rS}$ and a reconstructed real-world LR image, $\hat{\mathbf{I}}_{rL}$. To calculate adversarial loss, 4 discriminators are used; in the HR domain, $\mathcal{D}_{H1}$ discriminates between the real-world HR and synthesized SR images ($\mathbf{I}_{rH}$ and $\mathbf{I}_{sS}$), and $\mathcal{D}_{H2}$ discriminates between the real-world HR and real-world SR images ($\mathbf{I}_{rH}$ and $\mathbf{I}_{rS}$). In the LR domain, $\mathcal{D}_{L1}$ discriminates between the synthesized LR and real-world LR images ($\mathbf{I}_{sL}$ and $\mathbf{I}_{rL}$), while $\mathcal{D}_{L2}$ discriminates between the real-world LR and reconstructed real-world LR images ($\mathbf{I}_{rL}$ and $\hat{\mathbf{I}}_{rL}$). In addition to adversarial loss, SCGAN also computes pixel loss between the real-world HR/LR image and its degraded/restored counterpart, as well as cycle consistency loss between the real-world HR/LR image and its counterpart that has undergone a full degradation/restoration (or vice versa) cycle.

4.3.3 WIPA

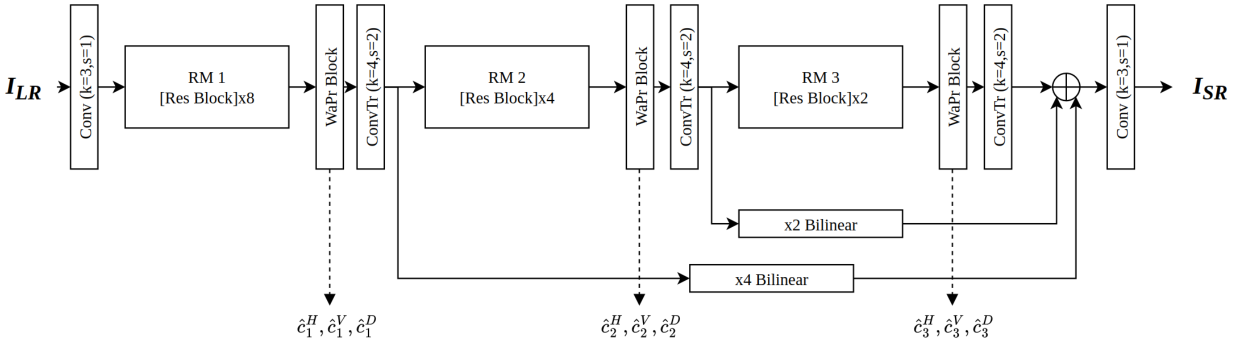


Figure 4.6: Architecture of the WIPA network.

The framework of WIPA (Dastmalchi & Aghaeinia, 2022) begins with a baseline SR network, made up of 3 Residual Modules with 8, 4, and 2 ResBlocks, respectively. The ResBlock architecture in WIPA differs from that used in SCGAN and is originally proposed by Ledig et al., 2017. Each Residual Module is followed by a 4×4 transposed convolution layer with a stride of 2 to increase the size of the intermediate feature maps. The output of these modules is resized to 128×128 using bilinear interpolation, and output through a 3×3 convolution layer with a stride of 1 and a Tanh activation function.

In order to integrate wavelet coefficient prediction into the baseline architecture, the authors propose a Wavelet Prediction block (WaPr Block) (Dastmalchi & Aghaeinia, 2022). The architecture of these blocks, as shown in Figure 4.7, consists of two convolution layers: the first is a 3×3 convolution layer, and the second is a 1×1 convolution layer. The 3×3 layer maps to the original 64-dimension wavelet-enriched feature maps, and the 1×1 layer reduces the number of wavelet-enriched channels down to 3 sub-bands to be matched with the ground-truth wavelet coefficients. The output of the WaPr Block consists of these

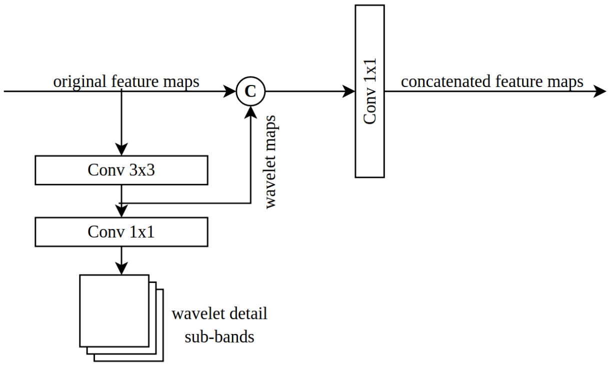


Figure 4.7: Architecture of the WaPr Blocks used by WIPA.

wavelet detail sub-bands concatenated to the original feature maps. The WaPr Blocks are placed just after the Residual Modules in the baseline architecture to form the full WIPA architecture, which is shown in Figure 4.6.

The WIPA framework contains 4 main loss functions. First, pixel-wise MSE loss is calculated between the ground-truth HR and generated SR images. The generated SR image is also passed into a discriminator network to calculate adversarial loss. To calculate the wavelet prediction loss, the ground-truth HR image is transformed into wavelet detail sub-bands at 3 different scales, and MAE is calculated between the ground-truth wavelet sub-bands and the sub-bands predicted by the WaPr Blocks of the network. Finally, two deep networks are used to calculate perceptual loss and identity loss. A VGG19 (Simonyan & Zisserman, 2015) network is used to calculate the perceptual loss between the HR and SR images, and SphereFace (W. Liu et al., 2017) is used to calculate the identity loss between the two images.

Table 4.1: Training parameters for the models used in the comparison study.

Model	Epochs	Batch Size	Learning Rate
DANv1	40,000	16	4e−4
SCGAN	200	64	1e−4
WIPA	200	32	1e−4

4.4 Comparison Study

To evaluate the performance of the selected models in this study, I follow the methodology outlined in the last chapter. All models were first trained on the FFHQ training split, following the training parameters shown in Table 4.1 for each model. After training, each model is evaluated on three testing datasets: CelebA with simple degradation, FFHQ with complex degradation, and WIDER FACE with real-world degradation. To evaluate the quality of the generated images, I use a variety of NR and FR IQA statistics from Chapter 2. To evaluate the quality of SR images with ground-truth images available (i.e., CelebA and FFHQ), I use both FR and NR statistics. PSNR and SSIM form the baseline metrics to approximate image quality with respect to the ground truth images, with LPIPS serving as a more complex, perceptual-based FR quality metric. I also use BRISQUE and NIQE as complementary NR quality metrics. To evaluate identity preservation, I use two metrics: face verification accuracy from ArcFace Deng et al., 2019 and magnitude scores from MagFace Meng et al., 2021. Since the WIDER FACE testing set has no corresponding ground-truth images against which to compare the SR results, I instead only use the NR statistics of BRISQUE, NIQE, and MagFace score to evaluate the quality of real-world SR. The results of this quality evaluation are discussed in the next chapter.

4.5 IP-SCGAN

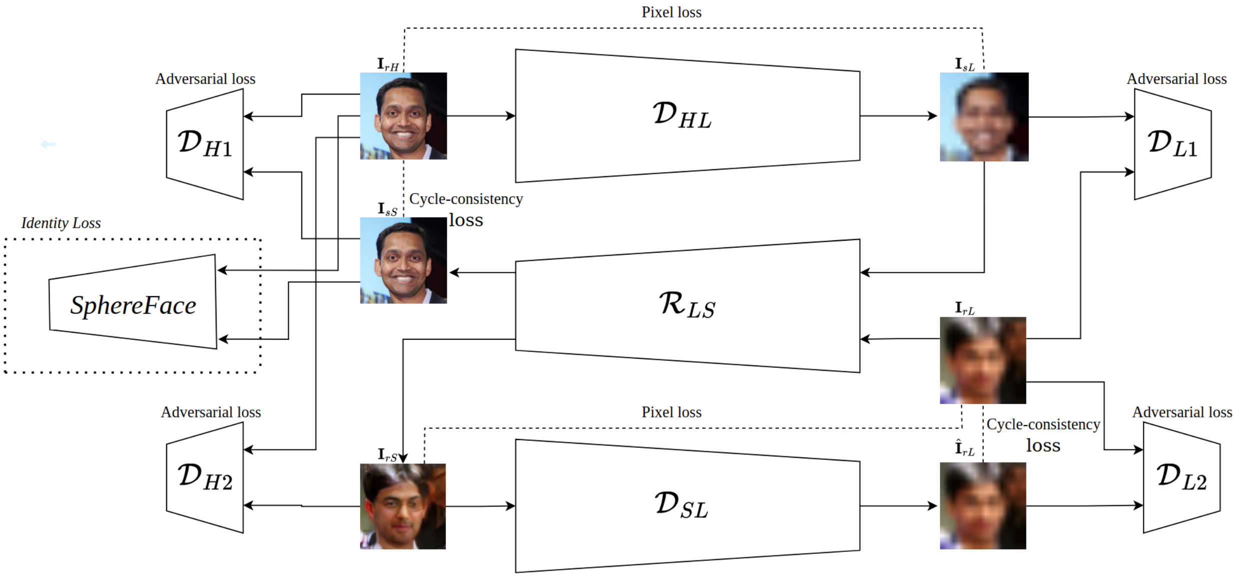


Figure 4.8: The loss framework for IP-SCGAN. The backbone framework is based on SCGAN (Hou et al., 2023); I place my contribution of a SphereFace network and identity loss in a dotted rectangle.

IP-SCGAN is a novel variant of the SCGAN framework, developed by me, that explicitly incorporates identity loss into the training loop. The main challenge in maintaining identity features in the SCGAN framework comes down to the fact that it is unsupervised and uses unpaired training data. However, by examining the training loop of SCGAN closely, it becomes apparent that the model itself creates a sort of "training pair" from the HR input image, namely $\mathbf{I}_{sL}$. When $\mathcal{R}_{LS}$ creates the synthesized SR image $\mathbf{I}_{sS}$, the following loss is calculated:

$$l_{\mathcal{R}_{LS}} = \theta l^{\mathbf{I}_{sS}}_{\mathcal{R}_{LS}} + \gamma l^{\mathbf{I}_{rS}}_{\mathcal{R}_{LS}} \tag{4.3}$$

where θ and γ are weighting parameters. The overall loss function for $\mathcal{R}_{LS}$ contains two terms: one for the loss of the synthetic SR image, and one for the loss of the real-world SR image. In turn, each of these losses are comprised of two loss terms, weighted with α and β:

$$l^{\mathbf{I}_{sS}}_{\mathcal{R}_{LS}} = \alpha l^{\mathcal{D}_{H1}}_{adv} + \beta l^{\mathbf{I}_{sS}}_{cyc}, \tag{4.4}$$

$$l^{\mathbf{I}_{rS}}_{\mathcal{R}_{LS}} = \alpha l^{\mathcal{D}_{H2}}_{adv} + \beta l^{\mathbf{I}_{rS}}_{pix}. \tag{4.5}$$

To increase the identity preservation capacity of SCGAN, I follow Dastmalchi and Aghaeinia, 2022 by using SphereFace to calculate the identity loss between $\mathbf{I}_{rH}$ and $\mathbf{I}_{sS}$. By weighting the identity loss with the term δ, I reformulate loss of $\mathcal{R}_{LS}$ with respect to $\mathbf{I}_{sS}$ as

$$l^{\mathbf{I}_{sS}}_{\mathcal{R}_{LS}} = \alpha l^{\mathcal{D}_{H1}}_{adv} + \beta l^{\mathbf{I}_{sS}}_{cyc} + \delta l^{\mathbf{I}_{sS}}_{id}. \tag{4.6}$$

To calculate identity loss, I use

$$l^{\mathbf{I}_{sS}}_{id} = 1 - S_C\big(Sphere(\mathbf{I}_{sS}),\ Sphere(\mathbf{I}_{rH})\big), \tag{4.7}$$

where S_C denotes cosine similarity and $Sphere$ represents the SphereFace-extracted feature map of a given image. The final loss framework for IP-SCGAN is shown in Figure 4.8.

Assuming that $\mathcal{D}_{HL}$ learns to approximate real-world degradation during model training, the way in which I implement identity loss allows IP-SCGAN to use the HR/LR pairs generated by this branch as a proxy for real-world training pairs. Thus, IP-SCGAN should be able to learn how to preserve identity features during image restoration while simultaneously learning to approximate real-world degradation. To test this hypothesis, I train and test IP-SCGAN on the same datasets as I use in the comparison study to evaluate its performance on SR and face recognition head-to-head against other state-of-the-art models. The results of these experiments are discussed in the next chapter.

4.5.1 Ablation Study

Table 4.2: The datasets used for each study in this thesis. Dataset names marked with an asterisk (*) indicate the training/testing subset was taken directly from the original SCGAN paper (Hou et al., 2023).

Experiment	Training	Simple	Complex	Real-World
SCGAN (Original)	FFHQ*	LFW*	FFHQ*	WIDER FACE*
Comparison Study	FFHQ	CelebA	FFHQ	WIDER FACE*
IP-SCGAN	FFHQ	CelebA	FFHQ	WIDER FACE*
IP-SCGAN Ablation	FFHQ*	CelebA	CelebA	N/A

To determine the best weighting for the identity loss term, I ran a simple ablation study. By holding α and β constant, I trained and tested the model with δ values in $[0.01,\ 0.05,\ 0.1,\ 1.0]$ to determine which value would be the best weighting to increase face recognition accuracy while maintaining competitive image quality statistics.

The datasets used in my ablation study differ slightly from the final evaluation of IP-SCGAN. In the ablation study, I compare the various δ values of my new loss function against the baseline SCGAN proposed in Hou et al., 2023. To this end, I decided to train my model on the same training segment of FFHQ as used in the original SCGAN paper, evaluating on my CelebA testing set under two different degradation conditions: simple and complex. In addition to the IQA metrics used in the comparison study and final evaluation of IP-SCGAN, I also use FID as a full-reference image quality metric.

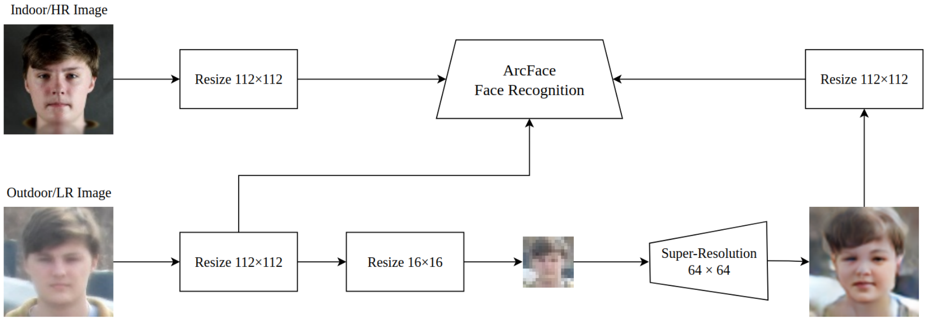


Figure 4.9: The framework for processing the images from MILAB-VTF(B) for super-resolution and face recognition.

4.5.2 Face Recognition on MILAB-VTF(B)

As a final test of real-world application for IP-SCGAN, I ran a face recognition experiment on the SR images produced by each model on the MILAB-VTF(B) (Bourlai et al., 2024) dataset. I take the 238 LR images that have been cropped and aligned with FFHQ, as well as the HR gallery images from the 46 subjects, and use ArcFace (Deng et al., 2019) to run face recognition on the query images. Fig. 4.9 shows the framework I use for processing the images for super-resolution and face recognition. After resizing all HR and LR images to a resolution of 112 to be compatible with ArcFace, I further downsample the LR images into a 16×16 resolution to prepare for SR. After super-resolving the resulting test images to a size of 64×64 using IP-SCGAN, SCGAN, WIPA[2], and bicubic upsampling, I upscale the resulting images to a resolution of 112×112 for face recognition. As a baseline, I compare the FR results from each super-resolution model against the original 300m outdoor images that have been downsampled directly to 112×112. This procedure allows me to compare the information loss between a native-resolution

[2] Following my procedure in the comparison study, images processed by WIPA are super-resolved to a resolution of 128×128 before being downsampled back to 64×64 for comparison with other methods.

real-world image (i.e., an image restoration task) and a real-world image that has to be upscaled to be used for face recognition (i.e., a super-resolution task). To evaluate the face recognition results, I record rank-1 and rank-5 accuracy, as well as a Receiver Operator Characteristic (ROC) curve and Area Under Curve (AUC) values.

Chapter 5

Results

In this chapter, I demonstrate the results of both of my studies discussed up until this point in the thesis. The first section discusses the results of the comparison study, and the second section discusses the SR results of IP-SCGAN.

5.1 Comparison Study

5.1.1 Comparison Study Results

The results of full-reference evaluation of the models used in the comparison study are shown in Table 5.2. The results of no-reference evaluation are shown in Table 5.1.

The first observation I make about the results concerns the performance of DANv1 (Luo et al., 2020), the only general SR algorithm evaluated during this study. Jiang et al., 2021 note that general-purpose SR methods often perform better than face-specific methods on FSR in terms of both PSNR and SSIM. The findings from this comparison reaffirm this observation, as DANv1 outperforms all other models on

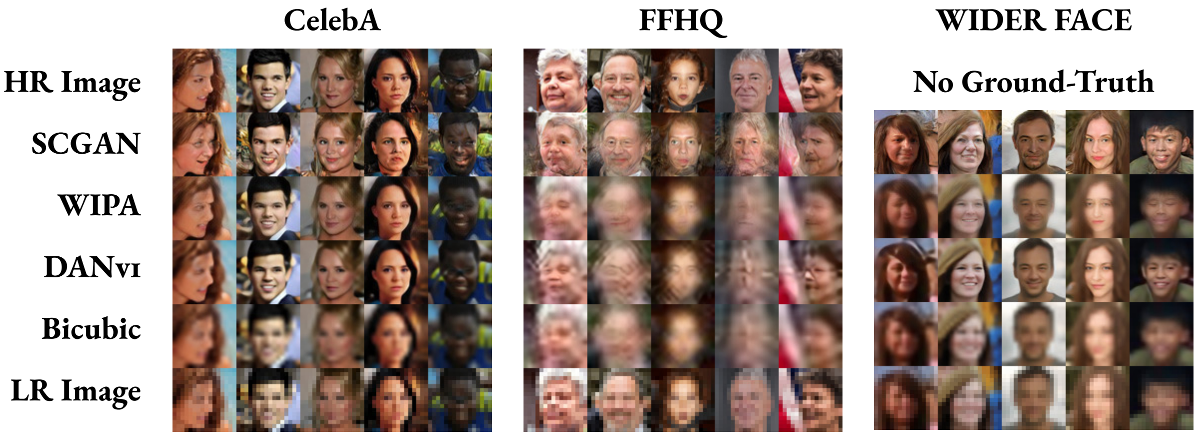


Figure 5.1: Comparison of visual quality of each SR method on CelebA, FFHQ, and WIDER FACE, with a scale factor of 4. Best viewed digitally; zoom in for a better view.

Table 5.1: Results of no-reference evaluation on super-resolution results with a scale factor of 4. Best values are shown in **red**, and second-best values are shown in *blue*.

Dataset	Degradation	Model	NIQE ↓	BRISQUE ↓	MagFace Score ↑
CelebA	Bicubic	WIPA	*8.13*	*37.76*	**23.41**
		SCGAN	**5.826**	**22.71**	*23.24*
		DANv1	11.11	45.93	22.90
		Bicubic	12.09	55.58	21.40
FFHQ	Complex	WIPA	*11.22*	*51.28*	*21.85*
		SCGAN	**5.29**	**17.76**	**22.41**
		DANv1	12.10	51.76	21.03
		Bicubic	12.91	58.90	21.74
WIDER FACE	Real-World	WIPA	*11.63*	50.50	*23.08*
		SCGAN	**5.61**	**21.36**	**23.32**
		DANv1	*11.63*	*46.87*	22.91
		Bicubic	11.94	56.13	21.91

these two metrics on the CelebA dataset; it also outperforms SCGAN on the FFHQ dataset on these metrics. However, when considering metrics that more closely mirror human perception (e.g., LPIPS, BRISQE, and NIQE), DANv1 barely performs better than baseline bicubic upsampling on both datasets.

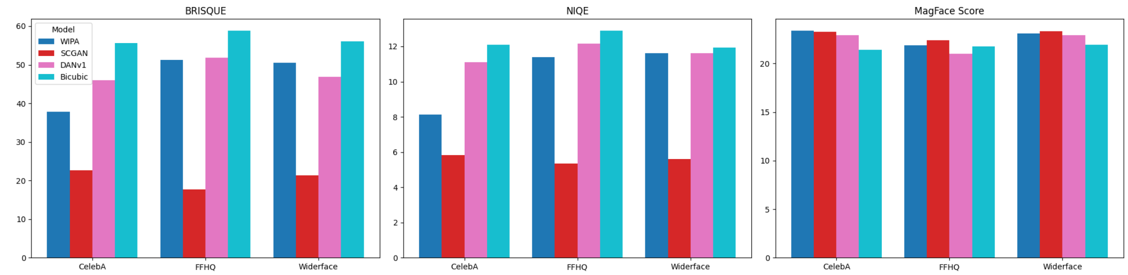


Figure 5.2: Graphs of results from no-reference evaluation on super-resolution with a scale factor of 4. Zoom in for a better view.

Table 5.2: Results of full-reference evaluation on super-resolution results with a scale factor of 4. Best values are shown in **red**, and second-best values are shown in *blue*.

Dataset	Degradation	Model	PSNR ↑	SSIM ↑	LPIPS ↓	Verif. Acc.
CelebA	Bicubic	WIPA	*27.10*	*0.8256*	**0.0595**	**86.56%**
		SCGAN	21.33	0.6351	*0.0740*	31.68%
		DANv1	**27.28**	**0.8392**	0.0960	*80.76%*
		Bicubic	23.98	0.6758	0.3740	13.44%
FFHQ	Complex	WIPA	**21.18**	**0.4800**	*0.3842*	*21.86%*
		SCGAN	19.37	0.4377	**0.1932**	**23.20%**
		DANv1	*21.00*	0.4701	0.3875	20.15%
		Bicubic	**21.18**	*0.4762*	0.5531	19.64%

This finding demonstrates that while general-purpose methods may outperform face-specific methods on PSNR and SSIM, these two metrics are not a reliable way to compare the resulting image quality of super-resolved face images between two models, since DANv1 evidently reconstructs information that is beneficial in terms of PSNR/SSIM, but not in terms of perceptual quality.

WIPA (Dastmalchi & Aghaeinia, 2022), which explicitly incorporates identity loss but fails to account for image degradation beyond bicubic downsampling, unsurprisingly boasts the highest ArcFace verification accuracy and MagFace score on the CelebA dataset. However, its performance rapidly declines on all metrics as image degradation intensifies. Conversely, SCGAN (Hou et al., 2023) performs the best on

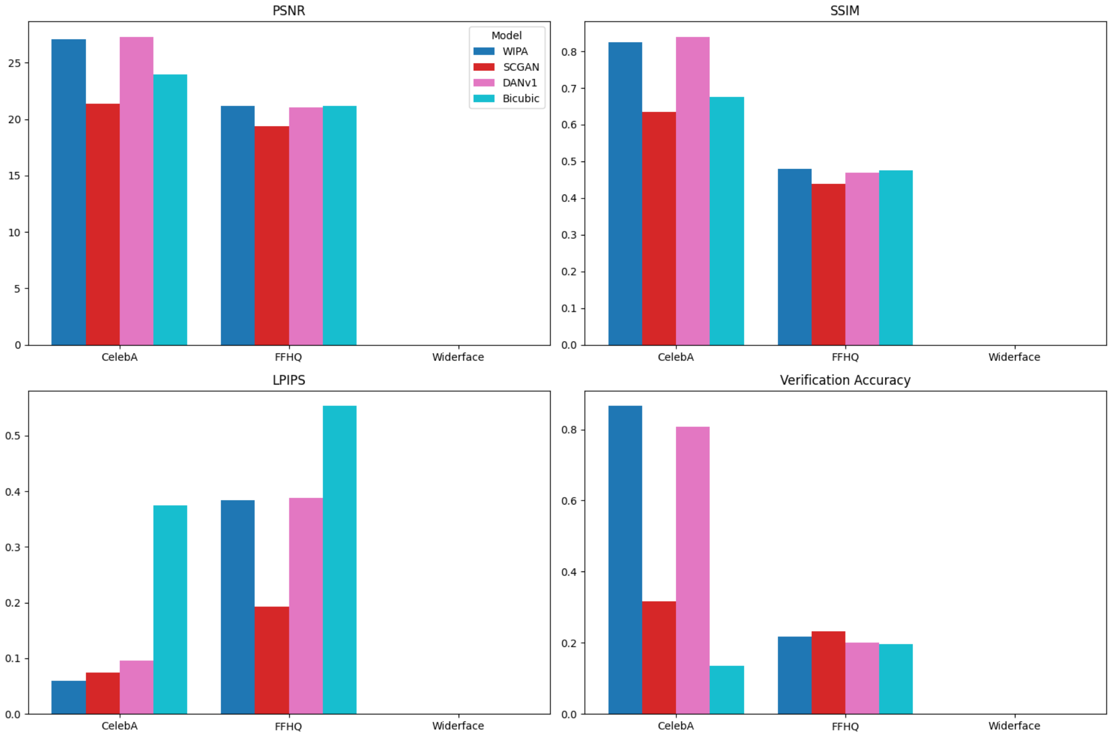


Figure 5.3: Graphs of results from full-reference evaluation on super-resolution with a scale factor of 4. Zoom in for a better view.

nearly all IQA metrics at complex and real-world degradation, especially in terms of NIQE and BRISQE, where it greatly outperforms all other approaches on all datasets. However, the results of the comparison study demonstrate that SCGAN's major weakness is identity preservation; while it does perform the best out of all models on FR for the FFHQ dataset, its verification accuracy is only 4 percentage points higher than bicubic upsampling. Moreover, it also only achieves a 32% verification accuracy on CelebA—by far the lowest of the three deep-learning based SR techniques.

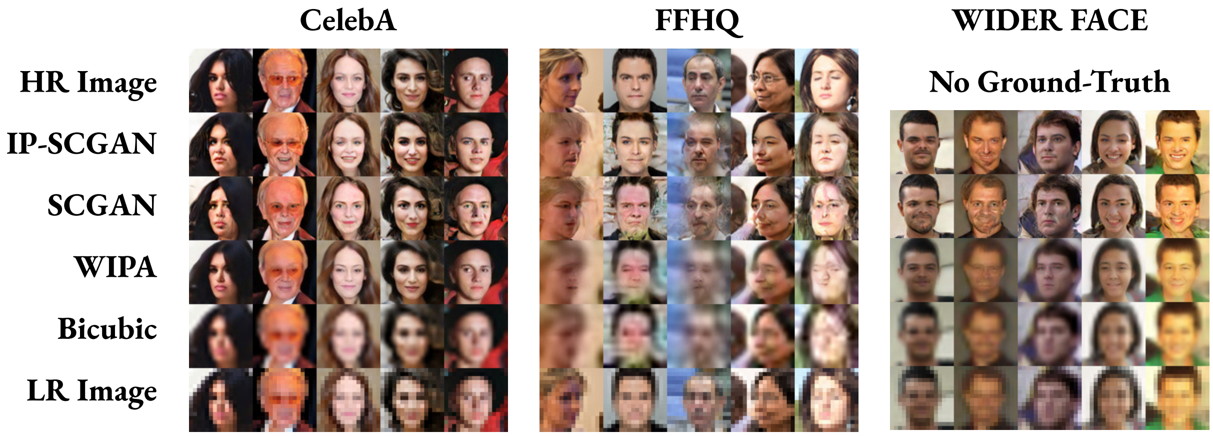


Figure 5.4: Comparison of visual quality of each SR method on CelebA, FFHQ, and WIDER FACE, with a scale factor of 4. Best viewed digitally; zoom in for a better view.

5.2 IP-SCGAN

5.2.1 IP-SCGAN Results

Table 5.3: Results of IQA and verification accuracy of the IP-SCGAN experiment on benchmark datasets. Best results are shown in **red**, second-best results are shown in *blue*.

Dataset	Model	PSNR	SSIM	LPIPS	BRISQUE	NIQE	Verif. Acc. (%)
CelebA	IP-SCGAN	21.52	0.6554	*0.0698*	*28.45*	*6.098*	*36.96*
	SCGAN	21.33	0.6351	0.0740	**22.71**	**5.826**	31.68
	WIPA	**27.10**	**0.8256**	**0.0595**	37.76	8.130	**86.56**
	Bicubic	*23.98*	*0.6758*	0.3740	55.58	12.09	13.44
FFHQ	IP-SCGAN	*20.05*	**0.4975**	**0.1632**	*23.15*	*6.455*	**28.36**
	SCGAN	19.37	0.4377	0.1932	**17.76**	**5.29**	*22.41*
	WIPA	**21.18**	*0.4800*	0.3842	51.28	11.22	21.85
	Bicubic	**21.18**	0.4762	0.5531	56.13	12.91	19.64
WIDER FACE	IP-SCGAN				*26.62*	*6.095*	
	SCGAN				**21.36**	**5.61**	
	WIPA				50.50	11.63	
	Bicubic				56.13	11.94	

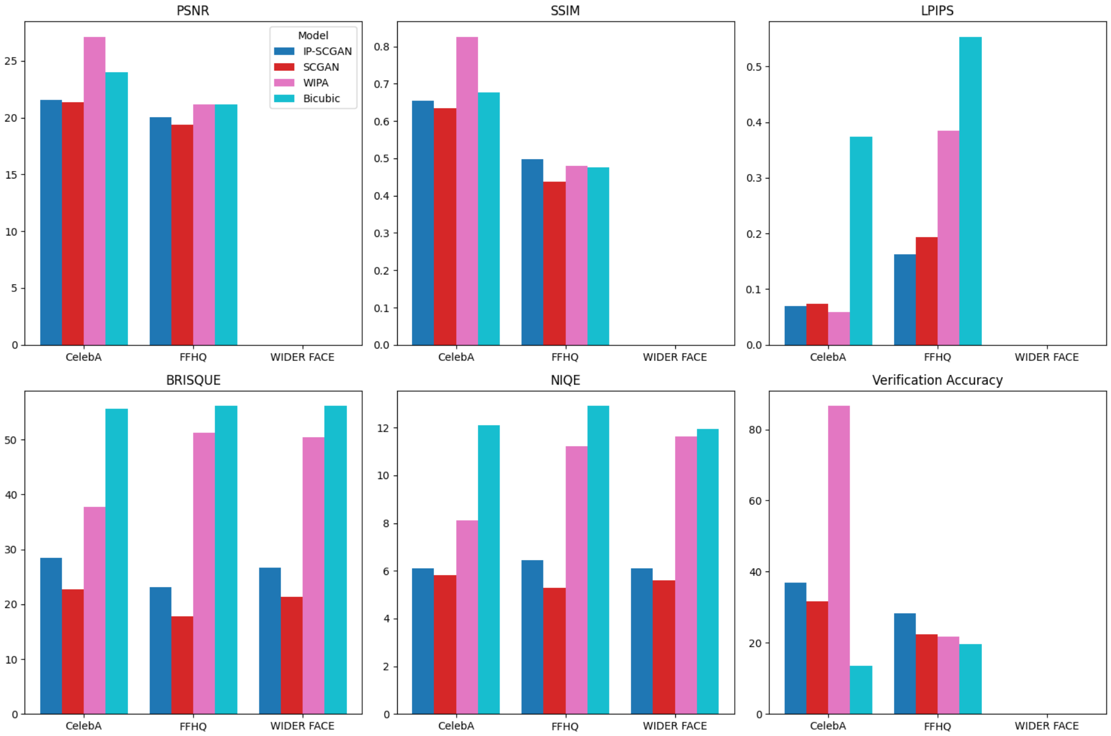


Figure 5.5: Graphical representation of the results from the IP-SCGAN study. Zoom in for a better view.

The results of evaluating the FSR performance of IP-SCGAN are shown in Table 5.3 and Fig. 5.5. I compare the performance of IP-SCGAN against the two models that inspired it, SCGAN and WIPA; all three models are also compared against baseline bicubic upsampling.

These results show that it is indeed possible to increase the identity preservation (face verification performance) of SCGAN by incorporating SphereFace identity loss. It is not surprising that WIPA maintains the best overall SR performance on bicubic downsampling, as this is the degradation environment for which the algorithm was designed and in which it demonstrates state-of-the-art performance (Dastmalchi & Aghaeinia, 2022) . However, IP-SCGAN outperforms baseline SCGAN in this setting on PSNR, SSIM, LPIPS, and verification accuracy.

Model performance on complex degradation (FFHQ) highlights the superiority of IP-SCGAN over the other models. IP-SCGAN outperforms both SCGAN and WIPA in terms of SSIM, LPIPS, and verification accuracy, and it far outperforms WIPA in both BRISQUE and NIQE at this degradation setting. These results demonstrate that IP-SCGAN retains the baseline SCGAN's ability to approximate complex and unknown degradation settings while simultaneously increasing identity feature preservation under aggressive image degradation.

IP-SCGAN Ablation Study

Table 5.4: Results for the ablation study of IP-SCGAN on CelebA. Super-resolution is performed with a scale factor of 4. Best restuls are shown in **red**, and second-best results are shown in *blue*.

Dataset	Model	FID	LPIPS	BRISQUE	NIQE	Verif. Acc. (%)
CelebA	Baseline	35.45	0.0756	28.98	**6.21**	32.88
	Delta_1.0	29.02	0.07498	*27.69*	6.665	34.16
	Delta_0.1	32.73	0.0870	30.20	7.208	27.02
	Delta_0.05	**25.37**	**0.0663**	**24.81**	*6.378*	*38.62*
	Delta_0.01	*26.95*	*0.0738*	34.50	7.349	**42.36**
CelebA_complex	Baseline	62.47	0.1597	22.30	**5.931**	16.94
	Delta_1.0	**49.93**	*0.1547*	23.01	*6.860*	**18.90**
	Delta_0.1	71.02	0.1703	*21.10*	7.359	15.70
	Delta_0.05	*58.39*	0.1583	**17.88**	6.869	*17.96*
	Delta_0.01	57.61	**0.1536**	29.72	7.579	17.60

The results of the IP-SCGAN ablation study are shown in Table 5.4. The purpose of the ablation study was to determine the best value of δ for weighting the identity loss term within the overall SCGAN loss framework. Results from each value of δ are compared against the baseline SCGAN model.

In general, δ values of 0.05 and 1.0 performed the best in the study. Surprisingly, decreasing δ all the way down to 0.01 actually performed the best on simple degradation in terms of face verification accuracy, but this impact was not similarly observed for complex degradation. In the end, I chose 0.05 to be the δ value, based on its superior performance on IQA and face verification scores under simple degradation, as well as its competitive IQA statistics and second-best face verification score under complex degradation.

5.2.2 Face Recognition Results on MILAB-VTF(B)

The Rank-1, Rank-5 and AUC values for the FR comparison are shown in Table 5.5. Fig. 5.6 shows the ROC curves for face recognition on each model, and Fig. 5.7 shows a comparison of the visual quality of super-resolved images for a selected number of subjects. In these figures, "Original" is used to denote the baseline of outdoor, 300m images that were downsampled directly to 112×112 resolution, while "LR" is used to denote the outdoor images that have been downsampled to 16×16. The HR images come from the indoor gallery dataset.

Table 5.5: Face recognition results on the MILAB-VTF(B) dataset. Baseline results are **underlined**. For SR models, best results are shown in **red**, second-best results are shown in *blue*.

Model	Rank-1 Acc (%)	Rank-5 Acc (%)	Area Under Curve (AUC)
Original	**47.06**	**48.74**	**0.35**
IP-SCGAN	**3.36**	*3.78*	0.22
SCGAN	*0.84*	1.26	0.30
WIPA	*0.84*	**4.62**	**0.47**
Bicubic	0.42	*3.78*	*0.38*

The results of these experiments were quite surprising, especially given the verification accuracy scores of the various SR algorithms on the complex LR images from FFHQ. If complex downsampling were truly a representative model of real-world degradation, one would expect to see higher Rank-1,5 accuracy scores from all super-resolution models. Instead, Rank-1,5 scores for all SR algorithms do not go higher

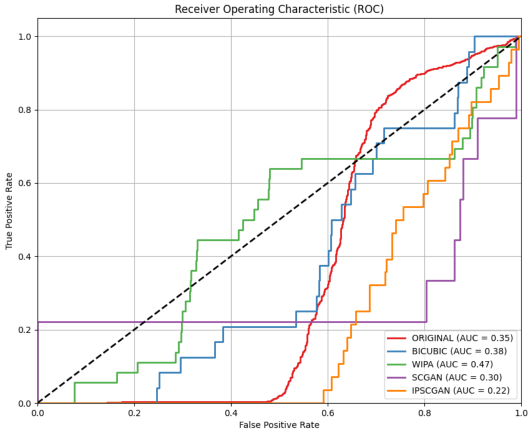


Figure 5.6: ROC curves for face recognition on the MILAB-VTF(B) dataset.

than 5%, indicating that model performance on face preservation under real-world degradation was largely inadequate for all SR models.

Even though FR accuracy from images created by all model was extremely poor, these results still indicate that adding identity loss to the SCGAN framework can increase identity preservation. In this experiment, images super-resolved by IP-SCGAN achieve a Rank-1 accuracy that is three times higher than those generated from SCGAN. Even though the accuracy score is still under 4%, the fact that it is significantly higher than the Rank-1 score achieved by SCGAN demonstrates the superiority of my novel technique over the original SCGAN model with respect to identity preservation at real-world degradation settings.

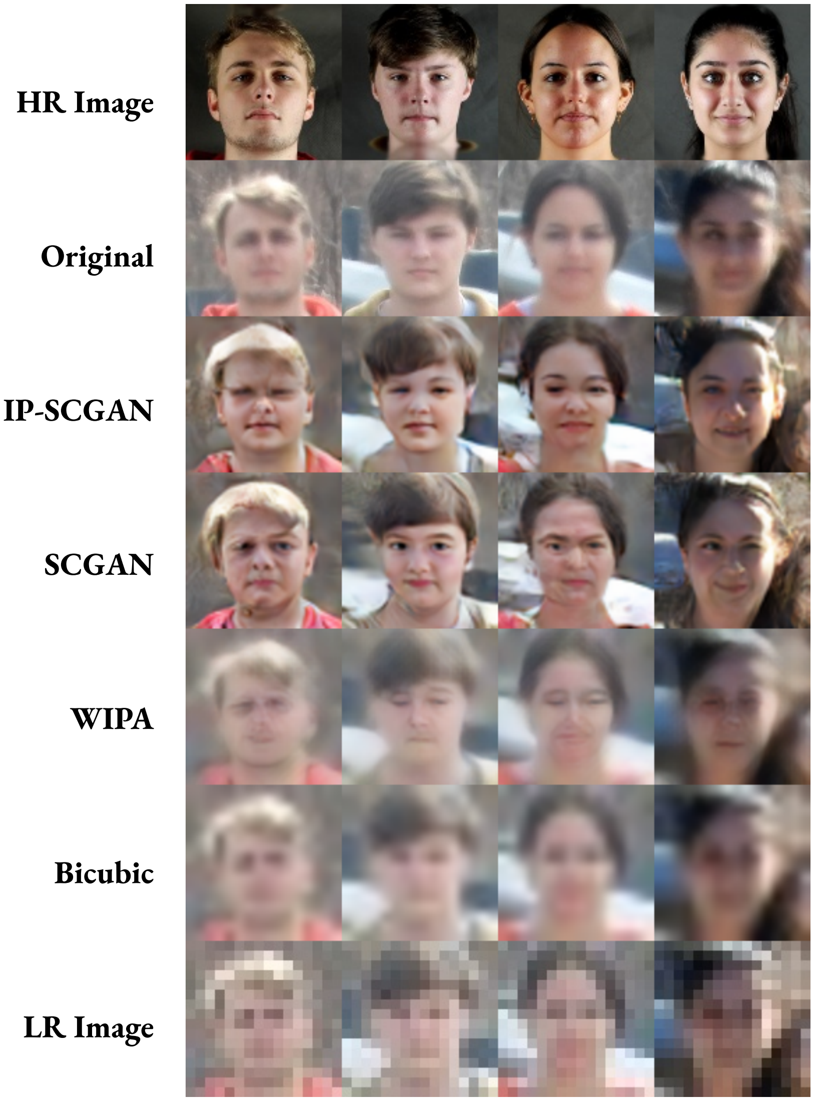


Figure 5.7: Comparison of visual quality on faces super-resolved from the MILAB-VTF(B) dataset with a scale factor of 4.

Chapter 6

Discussion

In this chapter, I provide a detailed theoretical discussion regarding the results of the experiments discussed in this thesis. I first discuss the results of the comparison Study, focusing on the trade-off between identity preservation and degradation estimation and the importance of image alignment in Face Super-Resolution. Next, I discuss the results of the IP-SCGAN study, focusing on its improvements over SCGAN and providing an explanation for the unsatisfactory results on the MILAB-VTF(B) dataset.

6.1 Comparison Study

In this section, I discuss the results of the comparison study of SR algorithms. To begin, I explore the theoretical and practical justifications behind the trade-off between identity preservation and degradation estimation. Then, I discuss the issue of image alignment during the FSR task, which was revealed to be a very important factor in overall FSR quality results.

6.1.1 Identity Preservation vs. Degradation Estimation

The results of the experiments discussed in this thesis have unveiled a concept that is crucial to understanding the state of current research on the topic of RWFSR, that being the domain gap. The substantial difference in data between simple and complex degradation greatly impacts model performance, and this trend can be seen throughout the results of these experiments. In the comparison study, WIPA performs extremely well on the simple bicubic downsampling it was designed to deal with; however, this performance degrades rapidly as soon as more complex degradation settings are introduced. SCGAN, conversely, handles complex and real-world degradation settings quite well, thanks to its domain transfer approach; however, the lack of a paired data paradigm in this case makes it very difficult for SCGAN to accurately preserve identity information, even under simple degradation.

When considering the performance of WIPA and SCGAN together, the trade-off between degradation complexity and identity preservation becomes clear: WIPA excels at identity preservation but fails at complex degradation settings, while the inverse is true for SCGAN. Even my attempt to merge these two algorithms in IP-SCGAN falls short of totally maintaining the strengths of each algorithm, as it could not achieve the same face verification accuracy as WIPA on simple degradation nor the IQA scores of SCGAN on complex degradation. IP-SCGAN is still a solid step forward, however, as it proves that it is possible to increase the identity preservation potential of the SCGAN framework while maintaining competitive IQA statistics on images with complex degradation.

The trade-off between identity preservation and complex degradation estimation has several potential causes. One is the nature of the RWFSR task as a multi-objective problem, balancing perceptual quality with identity preservation, so there is a Pareto Front of potentially optimal solutions. It is unlikely given the nature of multi-objective optimization problems that optimizing one objective will in turn optimize a different one; more commonly, there has to be compromise between the different objectives. Another reason we might observe this pattern has to do with the way researchers currently approach RWFSR when designing architectures and algorithms. As demonstrated during these studies, the best way to deal with one of these factors during model training is to explicitly account for it during model design and training. However, current research typically approaches the RWFSR problem from one of two angles: either explicitly considering identity, or explicitly considering complex degradation. Currently, there are very few approaches in the literature that even attempt to explicitly combine these two objectives.

There is also the issue of paired data: as stated in Chapter 2, it is often quite difficult or time-intensive to find identified, paired face data that contains both HR and real-world LR images of the same subject. Such datasets do exist, but typically are not benchmark datasets or publicly available. To train a hybrid method to both approximate real-world degradation and simultaneously learn identity preservation would require an extremely large version of such a dataset. Thus, it is no surprise that approaches like SCGAN that target real-world image degradation as their main objective instead rely on an unsupervised training framework, allowing them to train on unpaired HR and LR images from different benchmark datasets. However, the lack of training pairs greatly hinders these networks' ability to accurately learn to preserve identity information, as there is no way of "teaching" the model with data labels which faces are the same or different identities.

6.1.2 Alignment

One interesting observation I made during the conduction of my experiments concerns face alignment and its importance to the RWFSR task. As discussed by Jiang et al., 2021, the human face is a highly structured object with unique characteristics. For example, the relative positions of the eyes, nose and mouth of a frontal view of a human face are mostly consistent across all people. This feature of human faces makes it possible for certain algorithms to leverage facial landmark priors to assist in the FSR task, but it also creates limitations for certain algorithms.

One such limitation imposed on RWFSR algorithms due to the structured nature of the human face is the need for consistent alignment across training and testing image datasets. Consider as an example the SCGAN learning framework. As SCGAN is trained on unpaired HR and LR images, it learns the unique distribution of information in images in each of these domains. When trained on an HR dataset full of aligned images, it therefore has the potential to over-fit to the specific face feature landmark locations as they appear in the aligned dataset. Due to this over-fitting, the algorithm performs quite poorly on images that are not aligned with the original training set. Below is an example of this outcome; the third image has not been aligned in accordance with the FFHQ set, and has been super-resolved using SCGAN.

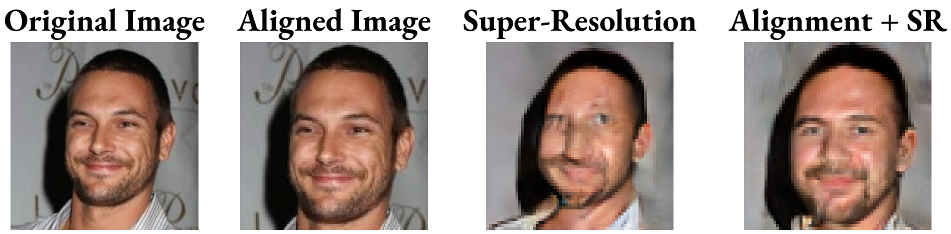


Figure 6.1: A visual comparison of images from CelebA, super-resolved by SCGAN, before and after alignment with the FFHQ dataset.

It is obvious that the resultant quality of the un-aligned image is worse than that of the FFHQ-aligned image. Observing these results, it is evidently clear that the algorithm is hallucinating face features in the un-aligned image in the location that it expects to find them based on FFHQ alignment; due to this, the resulting image does not have those facial features rendered accurately or placed appropriately inside the face area.

To determine the impact of face alignment on each SR algorithm evaluated in the comparison study, I tested each of the SR models on a version of my CelebA dataset that had not been aligned with FFHQ. Table 6.1 and Fig. 6.2 below show the results of this evaluation.

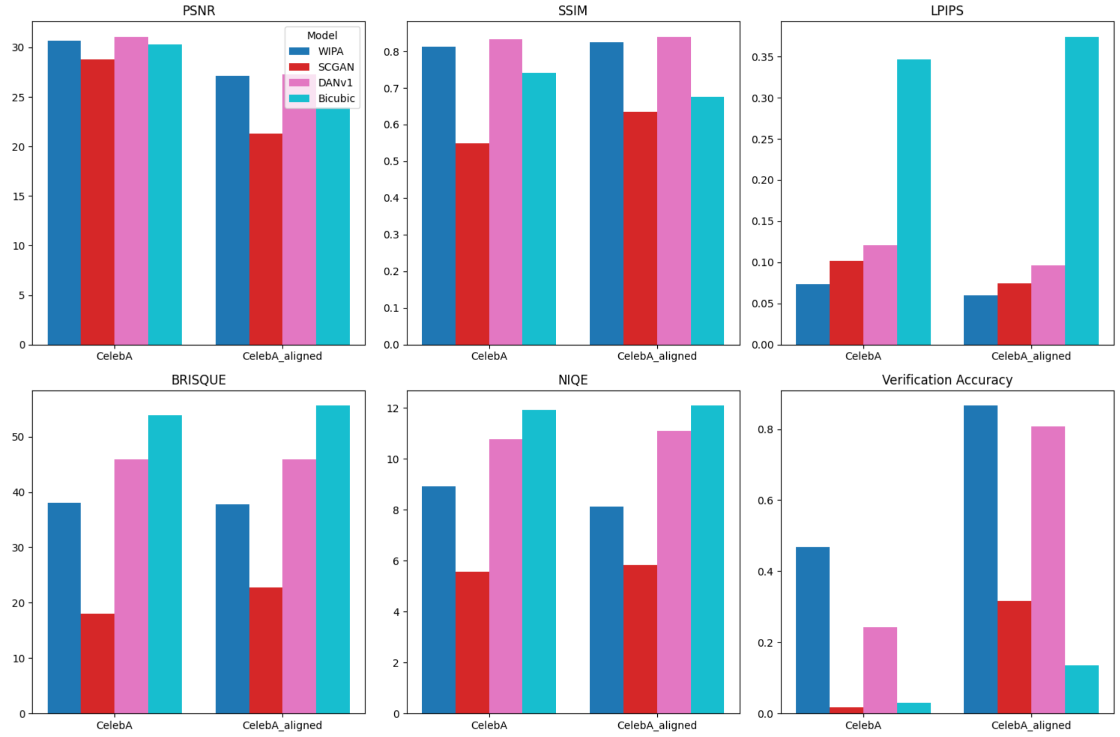


Figure 6.2: Graphs of results of SR + IQA evaluation on CelebA, before and after alignment with FFHQ.

Interestingly, face alignment did not have much impact on a majority of the IQA metrics on most models. Of the three models, SCGAN was impacted the most by face alignment, most likely due to its unsupervised learning structure. However, face alignment had a large impact on all models in terms of face verification accuracy, as evidenced by a multi-fold increase in face verification accuracy for all models between the un-aligned and aligned datasets. This investigation of face alignment is yet another example of the difference between image quality and identity preservation, as alignment largely did not significantly impact image quality from the models, but did substantially increase face verification accuracy for all models tested.

Table 6.1: Results of SR + IQA evaluation on CelebA, before and after face alignment with FFHQ. Best results are shown in **red**, second-best results are shown in *blue*.

Dataset	Model	PSNR	SSIM	LPIPS	Verif. Acc (%)	BRISQUE	NIQE
CelebA	WIPA	*30.71*	*0.8119*	**0.0730**	**46.78**	*38.01*	*8.92*
	SCGAN	28.80	0.5496	*0.1015*	1.82	**18.02**	**5.56**
	DANv1	**31.06**	**0.8335**	0.1207	*24.18*	45.92	10.78
	Bicubic	30.31	0.7416	0.3469	2.88	53.86	11.92
CelebA_aligned	WIPA	*27.10*	*0.8256*	**0.0595**	**86.56**	*37.76*	*8.13*
	SCGAN	21.33	0.6351	*0.0740*	31.68	**22.71**	**5.826**
	DANv1	**27.28**	**0.8392**	0.0960	*80.76*	45.93	11.11
	Bicubic	23.98	0.6758	0.3740	13.44	55.58	12.09

6.2 IP-SCGAN

In this section, I discuss the results of the IP-SCGAN algorithm experiments. First, I evaluate whether or not I was successful in improving identity preservation over SCGAN in complex and real-world degradation scenarios. Then, I explore the theoretical justifications for the overall poor results on super-resolving the MILAB-VTF(B) dataset with all models, before suggesting steps to improve model performance on the collected dataset moving forward.

6.2.1 Improvements Over SCGAN

The main goal of IP-SCGAN was to see if I could improve SCGAN's capacity to retain important identity information. Overall, I would say I was successful, if only marginally, in that goal. Looking at the results SR on the benchmark datasets, the most striking outcome of this experiment was that IP-SCGAN actually achieved the highest face verification accuracy on complex degradation. While its performance on face verification is eclipsed by that of WIPA on simple degradation, IP-SCGAN actually does not suffer the same relative drop in performance when moving from simple to complex degradation as WIPA. These results suggest that IP-SCGAN is a step closer to managing the trade-off between complexity of degradation and identity preservation than WIPA. IP-SCGAN also demonstrates superior performance to SCGAN on face verification across all degradation settings, indicating that the addition of identity loss to the SCGAN framework leads to increased identity preservation capabilities.

While IP-SCGAN improves the ability of the SCGAN framework to preserve identity information, the results of IQA evaluation on the benchmark datasets demonstrate that this improvement does not come at the cost of visual quality of the images; in fact, IP-SCGAN actually outperforms SCGAN on the FR metrics of PSNR, SSIM and LPIPS. In addition, its performance on the NR IQA metrics of BRISQUE and NIQE are still competitive with SCGAN, substantially outperforming both WIPA and baseline bicubic upsampling across all degradation settings.

The results of face recognition on the MILAB-VTF(B) dataset, which is the most accurate to a real-world deployment use case for RWFSR, again demonstrate that IP-SCGAN is able to significantly improve identity preservation over SCGAN. In terms of Rank-1 accuracy, IP-SCGAN achieves the highest face recognition score out of all SR models, over three times higher than that of either SCGAN or WIPA.

While IP-SCGAN achieved best-in-experiment Rank-1 accuracy on the MILAB-VTF(B) dataset, face recognition performance across all models on this datasets were highly dissapointing. In the next section, I explore some reasons why performing SR on images from the MILAB-VTF(B) datasets may have led to such poor face recognition results.

6.2.2 Face Recognition on MILAB-VTF(B)

One of the more interesting results of the experiments in this thesis concerns the face recognition results on the MILAB-VTF(B) dataset. While face verification accuracy scores among models vary from roughly 20% to 80% on benchmark datasets, Rank-1,5 accuracy scores on MILAB-VTF(B) never get above 5% for any super-resolution algorithm. In this section, I explore the reasons why FR accuracy scores on the MILAB-VTF(B) dataset are so low for these experiments.

I compare the results of my super-resolution evaluation with Philippe and Bourlai, 2024, who also work with augmenting the 300m outdoor images from the MILAB-VTF(B) dataset for face recognition. These image augmentation experiments produced face recognition results that are significantly better than those I achieved with super-resolution. Table 6.2 shows both the best results of face recognition using ArcFace on the MILAB-VTF(B) dataset after augmentation/super-resolution, as well as the results from performing FR on the baseline 300m images.

Table 6.2: A comparison of face identification metrics from an image restoration study and from my study, which focuses on image super-resolution.

Paper	Task	Result	Rank-1	Rank-5	AUC
Philippe and Bourlai, 2024	Restoration	Best	68.75	81.25	0.92
		300m Baseline	48.75	70.00	0.82
IP-SCGAN (Mine)	Super-Resolution	Best	3.36	4.62	0.47
		300m Baseline	47.06	48.74	0.35

There are a number of interesting observations to make about this comparison of results. The first is that in Philippe's paper, the FR results for the baseline 300m images vary from my own; while a Rank-1 accuracy of 47–48% is consistent across the two papers, the Rank-5 and AUC scores are much higher than what I record using ArcFace on the 300m baseline images. I attribute these inconsistencies to a difference in sample size between the two papers. While Philippe was able to leverage the entire MILAB-VTF(B) testing set for his experiments, my experiments worked with a substantially smaller subset of the testing set. Due to the alignment limitations of the SR algorithms I studied, I needed to use a face alignment script (Back, 2021) to align my testing images with FFHQ. Due to limitations in the face alignment script, and the fact that the 300m images are subject to real-world image degradation, I was not able to properly align and crop all the faces from all the subjects in the dataset. Because of this, it is highly likely that the smaller testing subset of MILAB-VTF(B) is improperly balanced or too small to provide useful Rank-5 and AUC values to my analysis.

Restoration vs Super-Resolution

The biggest reason Philippe's face recognition results score so much higher than mine is that although we work on the same datasets, the task at hand is actually quite different for the two papers. Philippe's paper focuses on *restoration*, where degradaded images are "cleaned up" by removing blur, noise and other degradations at the native resolution of the image. On the other hand, the *super-resolution* task focuses on performing these operations while simultaneously interpolating or generating new pixel values to increase the resolution of the image. Fig. 6.3 below highlights the difference between the two tasks. Shown in this figure is a sample image taken from the 300m outdoor testing set for MILAB-VTF(B) at a resolution of 112×112 pixels.

This figure demonstrates a low-quality input image generated from this face image for both a restoration task and a super-resolution task. On the left is the input image for a restoration algorithm, where the photo stays at its native resolution of 112×112. On the right is a rendering of the input image for

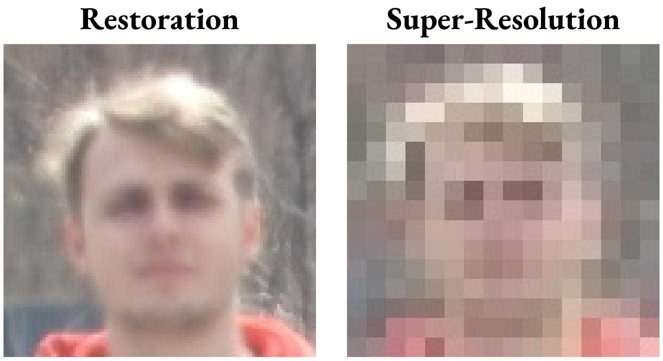


Figure 6.3: Examples of an input image for image restoration (left) and image super-resolution (right).

a super-resolution task at 16×16 resolution, rendered at 112×112 to showcase the information loss with respect to the original image.

The figure above makes it apparent that in the restoration task, the algorithm has much more pixel-level information to work with in order to deblur, denoise and restore the original image. In contrast, the super-resolution algorithm working with an image of size 16×16 has much less pixel-level information with which to generate a high quality image that restores the information lost in the original. Due to this inherent difference in these two tasks, it is significantly easier to generate high-quality face images that will be correctly identified by a FR system when restoring high-resolution, low-quality faces as opposed to low-resolution, low-quality faces.

Homogeneity of Data

Another factor that could account for the low SR performance on the MILAB-VTF(B) dataset is the over-fitting of models to the highly heterogeneous benchmark training sets. FFHQ (Karras et al., 2019) and WIDER FACE (Yang et al., 2016) are both heterogeneous datasets with large variation in pose, occlusion, lighting, and especially background. By comparison, the MILAB-VTF(B) dataset is much more constrained, featuring images of many different subjects but against a much more similar woodland background scene. Fig. 6.4 below shows a comparison of sample images from FFHQ and WIDER FACE, both of which are used to train SCGAN and IP-SCGAN, with images from MILAB-VTF(B).

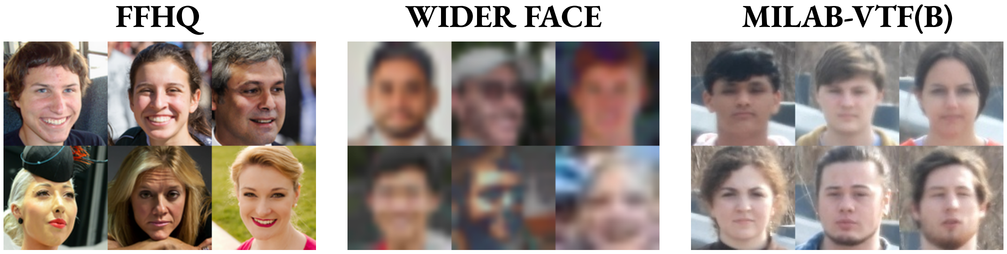
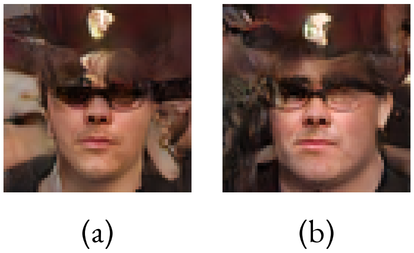


Figure 6.4: Sample images from FFHQ, WIDER FACE, and MILAB-VTF(B), showcasing the relative homogeneity of background scene in the latter dataset compared to the other two.

This comparison makes it apparent that there is much more intra-dataset variation in the benchmark datasets compared to the collected one. This makes intuitive sense, since the benchmark datasets are typically collected from large, publicly available sources such as Flickr, which was crawled to create FFHQ (Karras et al., 2019). On the other hand, the collected dataset was created specifically by the University of Georgia MILab using a single camera and only one or two imaging locations (Bourlai et al., 2024). Thus, a difference in intra-dataset variation is quite unsurprising between the two datasets.

However, the lack of background variation in the MILAB-VTF(B) dataset could have real consequences for RWFSR algorithms. Although a majority of the background of these images is cropped out during face alignment, part of the background is still visible; it is possible that FSR algorithms could over-fit to specific backgrounds. In this case, algorithms would learn to associate certain identities or facial features with certain features of the background (e.g., color), which would cause a mode collapse in the output of the model when evaluated on a more homogeneous dataset downstream. It is possible to observe this phenomenon when examining the output of the FSR models; for example, Fig. 6.5 below shows the output of IP-SCGAN on two different images from the WIDER FACE testing set.



(a) (b)

Figure 6.5: Two sample images of different subjects super-resolved from WIDER FACE using IP-SCGAN, showcasing the similarity in FSR results on images with similar backgrounds or non-face features.

The two images were taken from the same scene and feature similar features (e.g., the campaign hat), but are of two different identities. Nonetheless, the IP-SCGAN algorithm still seems to have hallucinated nearly identical features for the two images, including glasses and the shape of the nose. Without ground-truth images it can be difficult to tell if these features are in fact hallucinations or if they are actually accurate to the subjects in the images; however, it still serves as a solid example of how FSR algorithms may use features found outside of the actual face to influence the generated facial features.

One possible solution to this issue could be utilizing transfer learning when pulling an existing FSR algorithm into a new dataset or use case. By performing a majority of the training on benchmark datasets, the FSR algorithm can learn features and degradations that will be useful for the RWFSR task generally. Then, by fine-tuning the model on the final target dataset (which will likely be more homogeneous than the benchmarks), one can train the model on dataset-specific features, such as background scene or degradation created by a specific camera. By using this procedure, it will hopefully be possible to dissuade the FSR network from using non-face features like background as a heuristic for creating identity-matched SR images. This procedure could alleviate mode collapse, improve dataset-specific model performance, and increase downstream FR results from the RWFSR task.

Chapter 7

Conclusion

This thesis has explored the topic of Face Super-Resolution on images that have undergone complex synthetic and real-world image degradation, using two sets of experiments. The first set of experiments was a comparison study that trained and tested from scratch three state-of-the-art super-resolution models. As no model was built specifically to handle both of the main challenges facing RWFSR, namely restoration under real-world degradation and preservation of identity information, a comparison of these models was performed to determine which would make the most suitable baseline for a hybrid model. It was discovered during these experiments that models which train on paired image data were able to preserve identity information quite well under simple degradation settings. However, all models suffered and failed to produce identity-preserved images under complex and real-world degradation settings.

SCGAN, an unsupervised, domain transfer-based approach that learns FSR by way of unpaired training data, far outperformed the other two models in terms of image quality on complex and real-world degradation settings. However, it failed to create identity-preserved SR images even under simple degradation. Motivated by these strengths and weakness, I proposed a novel FSR model, called the Identity

Preserving, Semi-Cycled GAN (IP-SCGAN) to increase the identity preservation capacity of SCGAN. In order to do this, I added an identity loss term to the loss framework of SCGAN that is calculated using a SphereFace network.

To evaluate the proposed IP-SCGAN, I again performed a comparison experiment, this time evaluating the results of SCGAN and WIPA against that of IP-SCGAN on the benchmark datasets. I found that IP-SCGAN does indeed increase the identity preservation capacity of SCGAN, achieving the highest face verification accuracy on complex degradation of all models with IQA statistics that are competitive with SCGAN; however, it still does not produce identity-preserved SR images under simple degradation at the same rate as the supervised learning algorithms WIPA and DANv1.

As a final test of IP-SCGAN's ability to retain identity information, I evaluated it against SCGAN and WIPA on MILAB-VTF(B), a real-world face recognition dataset collected at the University of Georgia. I found that while IP-SCGAN again outperforms other models, this time achieving the highest Rank-1 accuracy out of all SR models, face recognition results across all models was extremely poor after running super-resolution with a scale of 4 on MILAB-VTF(B), especially in comparison to other papers that achieve much higher FR accuracy results after image restoration at a fixed resolution.

Overall, these experiments show that there is still much progress to be made toward a RWFSR model that can effectively handle both real-world degradation estimation and identity preservation of subjects in images. Future research may choose from multiple approaches to work towards solving this issue. First, there is a pressing need to address the weaknesses uncovered by this study and develop a hybrid model that explicitly accounts for both unknown degradation and face-specific identity features. Second, there is a need to collect a large, identified, "in-the-wild" dataset featuring images of the same subjects with and without real-world image degradation present. In the absence of such a dataset, training models to learn real-world degradation from “in-the-wild” images will pose in inherent limitation on their ability to preserve identity features, as the training images will likely need to be unpaired. Finally, there is a need to develop a model which handles independently image degradation estimation and subject identity preservation. While IP-SCGAN represents a step in the right direction towards this goal, the results from the comparison study demonstrated that other approaches to developing a hybrid model might be equally as viable, such as incorporating more sophisticated degradation estimation into a supervised, identity-preserving FSR framework like WIPA. Exploring these approaches will lead to better RWFSR models that can produce high-quality face images under unknown degradations while simultaneously assisting in downstream face recognition tasks.

Bibliography

Akansu, A., & Haddad, R. (2001). Multiresolution signal decomposition: Transforms, subbands, and wavelets. Elsevier Science. https://books.google.com/books?id=FokiyS75DDwC

Back, J. (2021). Ffhq-alignment. https://github.com/happy-jihye/FFHQ-Alignment

Bell-Kligler, S., Shocher, A., & Irani, M. (2019). Blind super-resolution kernel estimation using an internal-gan. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/5fd0b37cd7dbbb00f97ba6ce92bf5add-Paper.pdf

Bourlai, T., Rose, J., Mokalla, S. R., Zabin, A., Hornak, L., Nalty, C. B., Peri, N., Gleason, J., Castillo, C. D., Patel, V. M., & Chellappa, R. (2024). Data and algorithms for end-to-end thermal spectrum face verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *6*(1), 1–14. https://doi.org/10.1109/TBIOM.2023.3304999

Bulat, A., Yang, J., & Tzimiropoulos, G. (2018). To learn image super-resolution, use a gan to learn how to do image degradation first. *Proceedings of the European Conference on Computer Vision (ECCV).*

Chai, J., Zeng, H., Li, A., & Ngai, E. W. (2021). Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, *6*, 100134. https://doi.org/https://doi.org/10.1016/j.mlwa.2021.100134

Chen, H., He, X., Qing, L., Wu, Y., Ren, C., Sheriff, R. E., & Zhu, C. (2022). Real-world single image super-resolution: A brief review. *Information Fusion*, *79*, 124–145. https://doi.org/https://doi.org/10.1016/j.inffus.2021.09.005

Dastmalchi, H., & Aghaeinia, H. (2022). Super-resolution of very low-resolution face images with a wavelet integrated, identity preserving, adversarial network. *Signal Processing: Image Communication*, *107*, 116755. https://doi.org/https://doi.org/10.1016/j.image.2022.116755

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Dowson, D., & Landau, B. (1982). The fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, *12*(3), 450–455. https://doi.org/https://doi.org/10.1016/0047-259X(82)90077-X

El-Naggar, S., & Bourlai, T. (2019). Evaluation of deep learning models for ear recognition against image distortions. *2019 European Intelligence and Security Informatics Conference (EISIC)*, 85–93. https://doi.org/10.1109/EISIC49498.2019.9108870

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks.

Gu, J., Lu, H., Zuo, W., & Dong, C. (2019). Blind super-resolution with iterative kernel correction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Guo, G., & Zhang, N. (2019). A survey on deep learning based face recognition. *Computer Vision and Image Understanding*, *189*, 102805. https://doi.org/https://doi.org/10.1016/j.cviu.2019.102805

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2018). Gans trained by a two time-scale update rule converge to a local nash equilibrium. https://arxiv.org/abs/1706.08500

Hou, H., Xu, J., Hou, Y., Hu, X., Wei, B., & Shen, D. (2023). Semi-cycled generative adversarial networks for real-world face super-resolution. *IEEE Transactions on Image Processing*, *32*, 1184–1199. https://doi.org/10.1109/TIP.2023.3240845

Hsu, C.-C., Lin, C.-W., Su, W.-T., & Cheung, G. (2019). Sigan: Siamese generative adversarial network for identity-preserving face hallucination. *IEEE Transactions on Image Processing*, *28*(12), 6225–6236. https://doi.org/10.1109/TIP.2019.2924554

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132–7141. https://doi.org/10.1109/CVPR.2018.00745

Huang, H., He, R., Sun, Z., & Tan, T. (2017). Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Huang, H., He, R., Sun, Z., & Tan, T. (2019). Wavelet domain generative adversarial network for multi-scale face hallucination. *Int. J. Comput. Vision*, *127*(6–7), 763–784. https://doi.org/10.1007/s11263-019-01154-8

Jähne, B., Haussecker, H., & Geissler, P. (1999). *Handbook of computer vision and applications* (Vol. 2). Citeseer.

Jiang, J., Wang, C., Liu, X., & Ma, J. (2021). Deep learning-based face super-resolution: A survey. *ACM Comput. Surv.*, *55*(1). https://doi.org/10.1145/3485132

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks.

Kingma, D. P. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., & Shi, W. (2017). Photo-realistic single image super-resolution using a generative adversarial network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. *Proceedings of International Conference on Computer Vision (ICCV)*.

Luo, Z., Huang, Y., Li, S., Wang, L., & Tan, T. (2020). Unfolding the alternating optimization for blind super resolution. *Advances in Neural Information Processing Systems (NeurIPS)*, *33*.

Luo, Z., Huang, Y., Li, S., Wang, L., & Tan, T. (2021). End-to-end alternating optimization for blind super resolution.

Martin, M., & Bourlai, T. (2020). Unconstrained face recognition using cell phone devices: Faces in the wild. In T. Bourlai, P. Karampelas, & V. M. Patel (Eds.), *Securing social identity in mobile platforms: Technologies for security, privacy and identity management* (pp. 129–143). Springer International Publishing. https://doi.org/10.1007/978-3-030-39489-9_7

Meng, Q., Zhao, S., Huang, Z., & Zhou, F. (2021). MagFace: A universal representation for face recognition and quality assessment.

Mittal, A., Moorthy, A. K., & Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, *21*(12), 4695–4708. https://doi.org/10.1109/TIP.2012.2214050

Mittal, A., Soundararajan, R., & Bovik, A. C. (2013). Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, *20*(3), 209–212. https://doi.org/10.1109/LSP.2012.2227726

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks.

Nalty, C. B., Peri, N., Gleason, J., Castillo, C. D., Hu, S., Bourlai, T., & Chellappa, R. (2022). A brief survey on person recognition at a distance. https://arxiv.org/abs/2212.08969

Philippe, V., & Bourlai, T. (2024). Exploring image augmentation methods for long-distance face recognition using deep learning. *SoutheastCon 2024*, 1144–1150. https://doi.org/10.1109/SoutheastCon52093.2024.10500032

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. https://arxiv.org/abs/2112.10752

Shao, W.-Z., & Elad, M. (2015). Simple, accurate, and robust nonparametric blind super-resolution.

Shaoping Xu, S. J., & Min, W. (2017). No-reference/blind image quality assessment: A survey. *IETE Technical Review*, *34*(3), 223–245. https://doi.org/10.1080/02564602.2016.1151385

Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., & Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. https://arxiv.org/abs/1409.1556

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the inception architecture for computer vision. https://arxiv.org/abs/1512.00567

Szeliski, R. (2022). *Computer vision: Algorithms and applications*. Springer Nature.

Tsirikoglou, A., Eilertsen, G., & Unger, J. (2020). A survey of image synthesis methods for visual machine learning. *Computer Graphics Forum*, *39*(6), 426–451. https://doi.org/https://doi.org/10.1111/cgf.14047

Wang, X., Peng, J., Zhang, S., Chen, B., Wang, Y., & Guo, Y. (2022). A survey of face recognition. https://arxiv.org/abs/2212.13038

Wang, Z., & Bovik, A. (2002). A universal image quality index. *IEEE Signal Processing Letters*, *9*(3), 81–84. https://doi.org/10.1109/97.995823

Wang, Z., Bovik, A., Sheikh, H., & Simoncelli, E. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. https://doi.org/10.1109/TIP.2003.819861

Xiong, Y., Zhu, K., Lin, D., & Tang, X. (2015). Recognize complex events from static images by fusing deep channels. *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*.

Yang, S., Luo, P., Loy, C. C., & Tang, X. (2016). Wider face: A face detection benchmark. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Yuan, Y., Liu, S., Zhang, J., Zhang, Y., Dong, C., & Lin, L. (2018). Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric.

Zhang, Y., Liu, S., Dong, C., Zhang, X., & Yuan, Y. (2020). Multiple cycle-in-cycle generative adversarial networks for unsupervised image super-resolution. *IEEE Transactions on Image Processing*, *29*, 1101–1112. https://doi.org/10.1109/TIP.2019.2938347

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y. (2018). Residual dense network for image super-resolution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2020). Unpaired image-to-image translation using cycle-consistent adversarial networks.