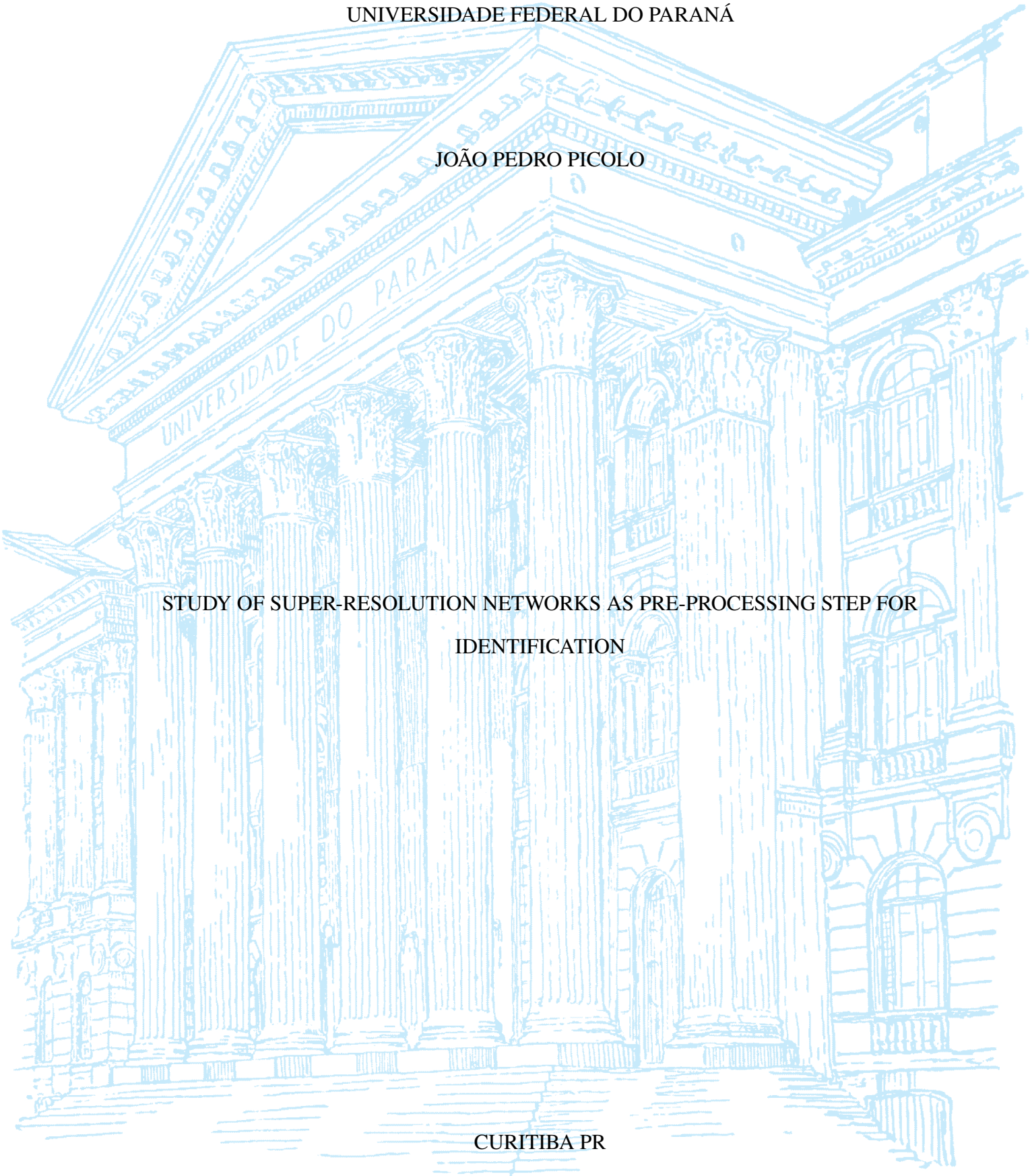
UNIVERSIDADE FEDERAL DO PARANÁ

JOÃO PEDRO PICOLO



STUDY OF SUPER-RESOLUTION NETWORKS AS PRE-PROCESSING STEP FOR IDENTIFICATION

CURITIBA PR

2023

JOÃO PEDRO PICOLO

STUDY OF SUPER-RESOLUTION NETWORKS AS PRE-PROCESSING STEP FOR IDENTIFICATION

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Ciência da Computação, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: David Menotti Gomes.

CURITIBA PR

2023

Universidade Federal do Paraná
Setor de Ciências Exatas
Curso de Ciência da Computação

Ata de Apresentação de Trabalho de Graduação II

Título do Trabalho: **STUDY OF SUPER-RESOLUTION NETWORKS AS PRE-PROCESSING STEP FOR IDENTIFICATION**

Autor(es):

GRR 20182659 None: JOÃO PEDRO PICOLO

GRR _____ Nome: _____

GRR _____ Nome: _____

Apresentação: Data: 13 / 12 / 2023 Hora: 09:00 Local: https://meet.google.com/xib-xrvw-tnq

Orientador: DAVID MENOTTI GOMES

Membro 1: RAFAEL OLIVEIRA RIBEIRO

Membro 2: MARCELO DOS SANTOS

(nome) (assinatura)

RAFAEL OLIVEIRA RIBEIRO:08989157757 Assinado de forma digital por RAFAEL OLIVEIRA RIBEIRO:08989157757 Dados: 2023.12.13 10:46:28 -03'00'

AVALIAÇÃO – Produto escrito		ORIENTADOR	MEMBRO 1	MEMBRO 2	MÉDIA
Conteúdo	(00-40)				40
Referência Bibliográfica	(00-10)				10
Formato	(00-05)				05
AVALIAÇÃO – Apresentação Oral					
Domínio do Assunto	(00-15)				15
Desenvolvimento do Assunto	(00-05)				05
Técnica de Apresentação	(00-03)				03
Uso do Tempo	(00-02)				02
AVALIAÇÃO – Desenvolvimento					
Nota do Orientador	(00-20)		**********	**********	20
NOTA FINAL		**********	**********	**********	100

Pesos indicados são sugestões.
Conforme decisão do colegiado do curso de Ciência da Computação, a entrega dos documentos comprobatório de trabalho de graduação 2 deve deve respeitar os seguintes procedimentos: Orientador deve abrir um processo no Sistema Eletrônico de Informações (SEI – UFPR); Selecionar o tipo: Graduação: Trabalho Conclusão de Curso; informar os interessados: nome do aluno e o nome do orientador; anexar esta ata escaneada e a versão final do pdf da monografia do aluno.; Tramita o processo para CCOMP (Coordenação Ciência da Computação).

RESUMO

Com a variedade de dispositivos disponíveis para a captura de imagens na atualidade, é comum que a qualidade das imagens obtidas varie de acordo com o sistema empregado. Com a finalidade de aprimorar a qualidade de imagens com baixa resolução, aplicam-se técnicas de super resolução de imagens. Estas técnicas revelam-se particularmente úteis em sistemas de câmeras de segurança, onde é crucial identificar indivíduos de interesse. Este trabalho propõe uma análise dos principais métodos de super resolução facial disponíveis publicamente, através de uma avaliação baseada na acurácia de um sistema de identificação, ao contrário de abordagens convencionais que se concentram em métricas de qualidade de imagem, como o índice de similaridade estrutural. Após a realização de testes estatísticos, conclui-se que muitos dos métodos disponíveis são eficazes em cenários específicos, especialmente para imagens altamente degradadas. No entanto, observa-se também que o método clássico de interpolação bicúbica para melhoria da qualidade apresenta resultados superiores a qualquer rede neural avaliada durante o desenvolvimento deste trabalho. Destaca-se como contribuição significativa a constatação de que, em cenários específicos, a aplicação de uma degradação bicúbica em uma imagem de baixa qualidade, seguida pelo aumento da resolução desta imagem para o tamanho original utilizando o método de interpolação pelos vizinhos mais próximos, pode ser uma etapa de pré-processamento valiosa na eliminação de ruído.

Palavras-chave: Processamento de imagens. Super resolução. Identificação.

ABSTRACT

The variety of devices available for capturing images nowadays often leads to variations in the quality of obtained images depending on the employed system. In order to enhance the quality of low-resolution images, image super-resolution techniques are commonly applied. These techniques prove to be particularly useful in surveillance camera systems, where it is crucial to identify individuals of interest. This work proposes an analysis of the main publicly available facial super-resolution methods through an evaluation based on the accuracy of an identification system, as opposed to conventional approaches that focus on image quality metrics, such as the structural similarity index. After conducting paired t-tests, it is concluded that many of the available methods are effective in specific scenarios, especially for highly degraded images. However, it is also observed that the classical method of bicubic interpolation for quality improvement yields superior results to any neural network evaluated during the development of this work. A significant contribution is highlighted in the finding that, in specific scenarios, applying bicubic degradation to a low-quality image, followed by resolution increase to the original size using the nearest neighbors interpolation method, can be a valuable preprocessing step in noise elimination.

Keywords: Image processing. Super-resolution. Identity Identification.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AC	Accuracy
CNN	Convolutional Neural Network
FSR	Face Super Resolution
GAN	Generative Adversarial Network
CGAN	Cyclic Generative Adversarial Network
HR	High Resolution
ISR	Image Super Resolution
LB	Lower Bound
LR	Low Resolution
MSE	Mean Squared Error
PSNR	Peak Signal-to-Noise Ratio
SOTA	State-of-the-art
SR	Super Resolution
SSIM	Structural Similarity Index
UB	Upper Bound
RNN	Recurrent Neural Network

CONTENTS

1 INTRODUCTION

Since the invention of the first camera in 1839, this technology experienced an explosion of options available for sale. Nowadays it is possible to find hundreds of different specifications for a camera such as the range of colors, the resolution, the sensor size, etc.

This explosion of options was accompanied by a large range of different noise types that can be present in an image. This noise was not a problem to be considered for a long time since the human brain has a good capacity to remove artifacts to detect objects, but with Industry 4.0 many of the recognition tasks were delegated to automatic recognition systems that do not have the same ability to remove noise as the human brain.

In particular, systems need to be able to ignore the noise and enhance quality to learn how to represent the objects presented in a particular image. This problem is known as Image Super Resolution (ISR). ISR has many applications in computer vision, such as the enhancement of satellite images allowing for better identification of items on the ground, enhancement of medical images to avoid misdiagnosis, enhancement of surveillance images for identity identification, etc. An example of a medical application can be seen in Fig. 1.1.

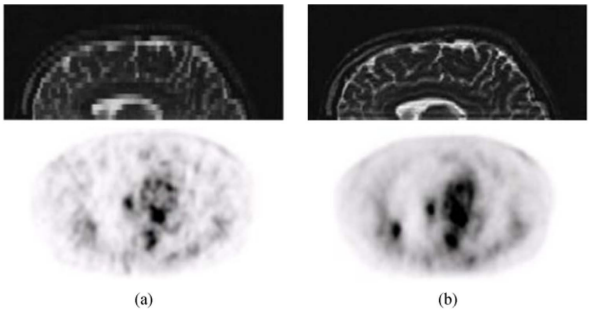


Figure 1.1: Example of a medical application. The first column has the original images, and the second column has the enhanced images. Image extracted from Yue et al. (2016).

State-of-the-art (SOTA) techniques for ISR, like the ones proposed by Liang et al. (2021) and Zhang et al. (2022a), explore multiple approaches for this task such as the use of Convolutional Neural Networks (CNN), Generative Adversarial Networks (GAN), Transformers, etc. It also exploited prior information from the subject to be reconstructed, such as landmarks for facial enhancement.

Other works also proposed the task of Video Super-Resolution for the recovery of a sequence of frames instead of a single image. This task is explored by Zhang et al. (2022b), in general, the result obtained is similar to the ones obtained for single images. Due to computational cost constraints, this work focuses on ISR performed over static images instead of using videos.

1.1 MOTIVATION

ISR techniques are particularly useful for enhancing the quality of images captured in surveillance systems, once that having a high-quality image is helpful when detecting criminals. Especially, it is important to enhance any faces presented in a surveillance camera as a way to identify the individual of interest. Some examples of this scenario can be seen in Fig. 1.2.

While most ISR works focus on enhancing the images with a pleasant human perspective, it is possible to observe that a minority of works consider practical applications of face recovery

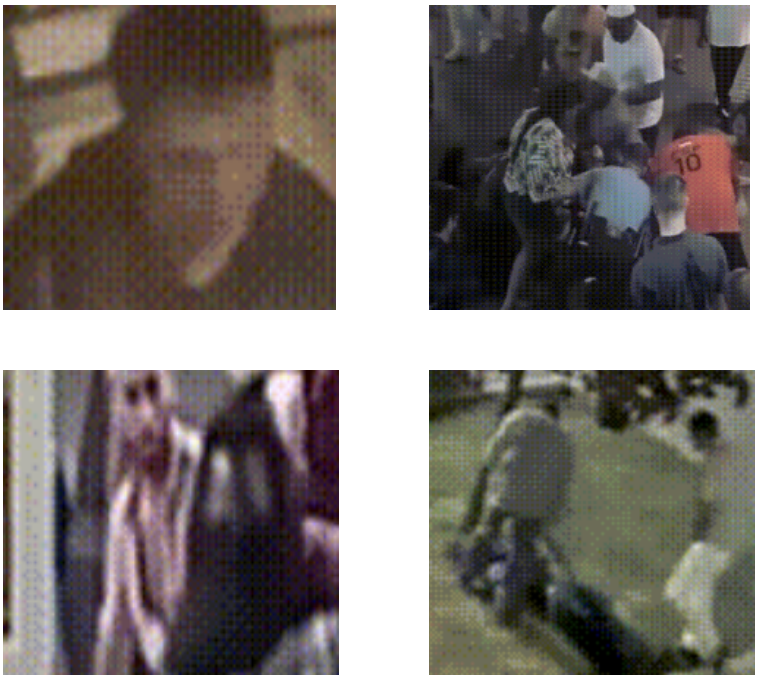


Figure 1.2: Example of suspects on surveillance systems. Images extracted from the RWF2000 dataset.

while performing experiments. The works that do focus on applications, like the ability to identify an identity from a recovered image, generally fit in one of these two cases: the used datasets have been manually downgraded to reproduce as much as possible surveillance camera systems or do not give much attention to these tests, performing them only over a subset of the whole real-world dataset.

This work proposes a more extensive evaluation of the super-resolution networks for identity identification by avoiding the problems mentioned above: SOTA networks are selected to recover faces from real-world datasets that have unknown degradation models and variations, once these faces are recovered they are fed to a identification network able to point out if the recovery was successful or not.

1.2 CHALLENGES

Besides the classical challenges related to posing, resolution, occlusion, and lighting variations presented in the datasets, a particularly difficult challenge when building an ISR technique for a surveillance system is to obtain data that represents a real-world scenario since it is difficult to artificially reproduce the noise generated by the different variations of cameras and sensors used in these systems. To address this problem, some works have been proposed on learning image degradation by using neural networks and others have been proposing bigger surveillance system datasets.

1.3 HYPOTHESES

During the development of this work, the hypothesis to be evaluated is that using super-resolution networks on low-quality images as a pre-processing step helps to improve identification accuracy. This is done by using a statistical approach called Paired Student's t-test which compares the means between two populations.

1.4 PROPOSED APPROACH

This work evaluates the use of ISR techniques for facial recovery using identification accuracy as the leading metric, as most of the existing works focus on metrics related to the quality of human perception.

1.5 CONTRIBUTIONS

Differently from previous works, a full accuracy evaluation performed over real-world facial datasets on ISR scenarios is performed. Through this evaluation, it was possible to observe that using the ISR methods on images with high amounts of degradation can be helpful for identification, but the same is not true for real-world datasets where the cameras already capture the image with high-quality.

Furthermore, classical approaches such as the bicubic interpolation upsampling and simple image enhancements showed better results when compared to the proposed ISR methods.

1.6 OUTLINE

The following chapters are divided as follows: Chapter 2 gives an overview of the theoretical background needed, bringing information about the Face Super-Resolution task, the main degradation methods applied over images, and the main assessment metrics used to evaluate the existing literature. Chapter 3 presents the existing literature on the task, by reviewing the main datasets used by the existing works, as well as by bringing the main approaches used to solve the proposed problem. Chapter 4 introduces the proposed approach of this work by going deeper into the description of the ISR and Identification networks used, as well as by doing a more extensive description of the used datasets. Chapter 5 describes the main experiments performed by describing the pre-processing steps and the test scenarios used, and also presents the results obtained for each result by bringing a final discussion over them. Finally, Chapter 6 will present the conclusion of this work, by including any limitations found during the process and suggesting future works.

2 THEORETICAL BACKGROUND

In this chapter, elementary concepts inherent to the studied field are presented. Section 2.1 explains the idea of Face Super Resolution and the different types of classification a technique can be put into. Section 2.2 analyzes the different methods used for artificially degrading images when building datasets. Section 2.3 presents an overview of the commonly used assessment metrics in this field. Section 2.4 elaborates a brief conclusion of the analysis made through this chapter.

2.1 FACE SUPER RESOLUTION

Face Super Resolution (FSR) consists of reconstructing High Resolution (HR) facial images from Low Resolution (LR) facial images, the reconstructed images are called Super Resolution (SR) images. This task was first introduced by Baker and Kanade (2000) as a domain-specific approach to the ISR problem, once it allows the use of prior knowledge when recovering face details instead of making assumptions about the structure of the image to be reconstructed. In Fig. 2.1 it is possible to visualize LR input images being reconstructed into SR images.

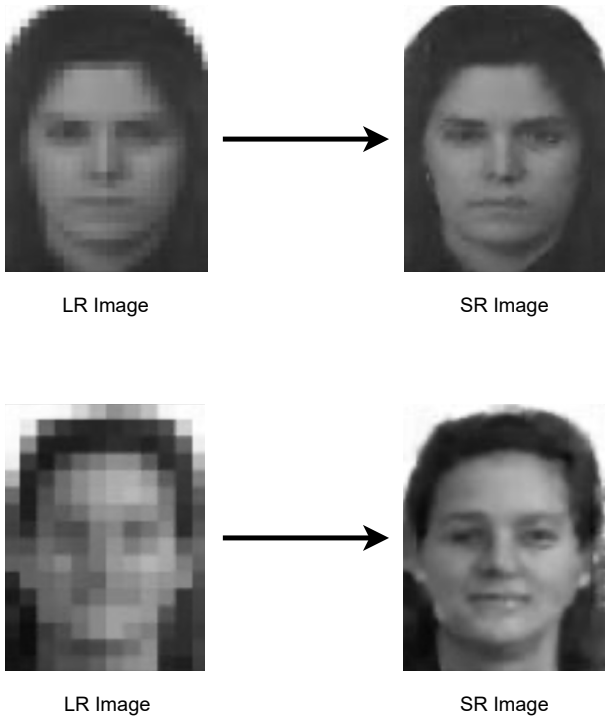


Figure 2.1: Example of images reconstructed using Super Resolution. Images extracted from Baker and Kanade (2000).

This work classifies the existing face reconstruction approaches in either Statistical or Deep Learning. The Statistical approaches use mathematical methods to manipulate a single LR input until an SR image is obtained, some examples are the use of Gaussian Pyramids proposed by Baker and Kanade (2000), Gaussian Distributions and Markov Networks proposed by Liu et al. (2001), etc.

Following a similar taxonomy to the one proposed by Jiang et al. (2021), the Deep Learning approaches are then broken down into one of the following categories: General, Prior-guided, Identity-preserving, or Reference. A summary of the descriptions provided by Jiang et al. (2021) is used to explain each method for the mentioned categories.

The most straightforward implementation of the Deep Learning methods is based on using General Networks to generate the SR image. These networks can be broken down into the following methods:

- **Global CNN methods** can learn how to represent the facial structure as a whole.
- **Local CNN methods** focus on recovering smaller face details such as contours.
- **Mixed CNN methods** fuse Global and Local methods to learn the facial structure while being able to recover facial details. This approach leads to better results once it gets more information than the methods above.
- **General GAN methods** use a base GAN structure consisting of two subnetworks: a generator responsible for recovering the SR image and a discriminator to distinguish the generated image from the HR image. These methods rely on artificial LR and HR image pairs generated by a known degradation, such as downsampling.
- **Generative GAN methods** approach the super-resolution problem with a different perspective, where the reconstruction problem is faced as a generation problem. This method uses a pre-trained GAN to generate SR that is similar to the LR images when downsampled.
- **Reinforcement RNN methods** ignore the contextual dependencies between different facial structures while learning the mapping function between the LR input and the SR.
- **Ensemble methods** are based on the usage of the three types of networks cited in the above methods, where each network generates an SR candidate. At the end of the process, all candidates are fused to generate the final SR image.

Since the FSR is a domain-specific problem of ISR, prior knowledge can be used to obtain better results. This fact led to the development of Prior-guided Networks, these methods extract facial information to assist the reconstruction process and can be divided into:

- **Pre-prior methods** initially extract the facial information from the LR input and then feed it to the beginning of the reconstruction network.
- **Parallel-prior methods** exploit the correlation between the extraction and the reconstruction tasks. The networks for each of these tasks are jointly trained and use an HR image to calculate prior-based loss.
- **In-prior methods** are based on the assumption that prior information extracted from the LR input is not a good representative of the necessary information. These methods then use a coarse process, like upsampling, to generate intermediate SR images. Once the prior information is extracted from the intermediate image, both these pieces of information are fed to the reconstruction network.
- **Post-prior methods** extract the facial information from the generated SR image. These pieces of information are then compared to the HR images to compute the prior-based loss that is then applied to the reconstruction network.

An important prior information that should not be ignored while reconstructing a facial image is the identity of the subject. Based on this assumption it was developed Identity-preserving Networks that can be divided into the following methods:

- **Face Recognition methods** tend to use a reconstruction network to generate the SR images that are subsequently fed to a recognition network that generates the identity loss between the SR input and the HR image. This loss is fed to the initial network to improve sequential results.
- **Pairwise methods** are based on the assumption that it is easier to obtain weakly labeled datasets. These networks generate a pair of SR at the same time for different LR inputs, then it is applied to the network an identity-preserving contrastive loss that is minimized when the pairs belong to the same identity and maximized otherwise.

All the previous categories only used the input LR images and their information to obtain the SR result. Based on the assumption that in some scenarios it is possible to have HR images of the same identity as the LR input, the Reference Networks exploit this information to boost the generated results. These networks can be split into the following methods:

- **Single-face methods** initially use the HR image and the LR input and feed both images to the construction network. The reconstruction networks can take advantage of multiple information such as landmarks, illumination, etc.
- **Multi-face methods** work similarly to the previous methods, the difference is based on the idea that since multiple HR images are available for the same identity, it is possible to assign different weights to each one of them during selection and training.
- **Dictionary methods** are based on the assumption that the set of HR images does not need to belong to the same identity as the LR input, it is only necessary to use a component dictionary to connect both pieces of information. These networks then explore similar facial structures from both sides.

2.2 DEGRADATION METHODS

A common challenge faced when building FSR networks is to find data on which the network can be trained. Since many of the available datasets do not have inherent LR images, it is necessary to use artificial degradation methods to generate LR images from an HR input. The degradation process can be formulated as follows

$$LR = f(HR, \theta), \tag{2.1}$$

where f represents a function that receives an HR image, θ degradation parameters and outputs an LR image. Multiple methods can represent the function f and the most common are the Gaussian Noise and Bicubic Interpolation.

Gaussian Noise is based on the generation of random numbers in a matrix following a normal distribution. The distribution is defined by the probability density function defined in Eq. 2.2, and is used to estimate the noise present in an image as described by Wu and Chang (2012).

$$p(z) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(z-\mu)^2}{2\sigma^2}} \tag{2.2}$$

where z represents a random Gaussian variable, μ is the mean of z, and σ is the standard deviation used to control the shape of the distribution. The generated matrix has the same size as the input HR image and is summed to it to output the noised image. This process can be seen in Fig. 2.2.

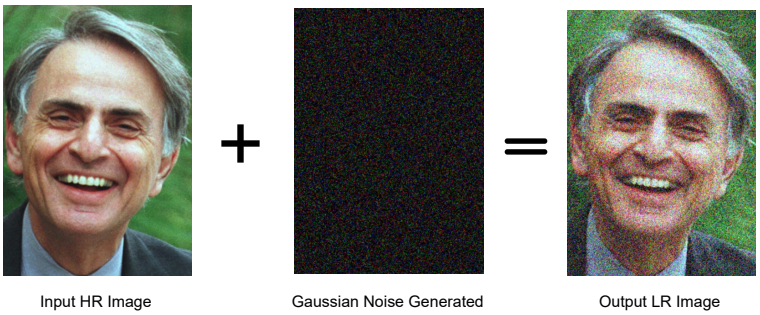


Figure 2.2: Example of Gaussian Noise applied to an image.

Bicubic Interpolation was proposed by Keys (1981), it is a mathematical algorithm applied over an image to estimate neighbor pixel values during resizing. The kernel function of this algorithm is defined in Eq. 2.3.

$$u(s) = \begin{cases} (a+2)|s|^3 - (a+3)|s|^2 + 1 & 0 \leq |s| < 1 \\ a|s|^3 - 5a|s|^2 + 8a|s| - 4a & 1 \leq |s| < 2 \\ 0 & 2 \leq |s| \end{cases} \tag{2.3}$$

where a is a predefined coefficient and s is a discrete value to be converted to a continuous value.

Let $t(x, y)$ be the pixel mapped on the resized image, and let I be a matrix containing the interval $[x-4, y-4] \times [x-1, y-1]$ of the input image. Then $t(x, y)$ can be expressed in Eq. 2.4.

$$t(x, y) = X \cdot I \cdot Y \tag{2.4}$$

where $\cdot$ represents the dot product of two matrices, X and Y are matrices defined by applying the kernel function to four predefined coordinate values of the nearest pixels in the input image.

An example of this method's application can be seen in Fig. 2.3.

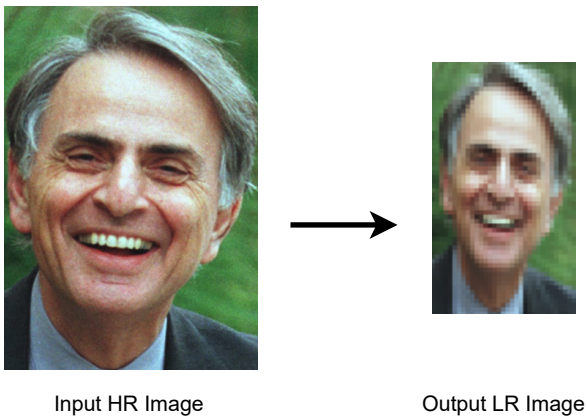


Figure 2.3: Example of Bicubic Interpolation applied to an image.

Even with the application of artificial techniques for this purpose, it is still not possible to reproduce real-world degradation synthetically since in the wild LR images have a wide impact

from type compression, the quality of the camera's sensor, and many other factors. To address this problem works such as the one proposed by Li et al. (2023) use CNNs to learn how to mimic the necessary degradation from images available in non-facial datasets.

2.3 ASSESSMENT METRICS

In computer vision, to understand if a network is outputting good results, both qualitative and quantitative analysis can be made. In this section, three main assessment metrics will be presented: Accuracy (AC), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM).

AC is used to evaluate classification models, in the context of FSR, this metric can be used when the work proposes to study the ability of a reconstruction model to generate an SR image that is good enough to find a match in the testing set. The AC metric is defined in Eq. 2.5.

$$AC = \frac{Correct\ \ Predictions}{Total\ \ Predictions} \tag{2.5}$$

It is important to notice that this metric not only depends on the reconstruction model but also on a separate face detection network so it does not necessarily represents the quality of the reconstructed images. AC also better represents identification scenarios rather than verification, this happens because in verification scenarios the risk of detecting false positives or false negatives may be higher once every image will be tested against the remaining images of the testing set.

PSNR uses the Mean Squared Error (MSE) to evaluate how similar two images are, the MSE function is defined in Eq. 2.6.

$$MSE(SR, HR) = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} |LR(i, j) - HR(i, j)|^2 \tag{2.6}$$

where SR and HR are images of size M x N. Since the MSE calculates the difference between each pixel of the compared image, the PSNR metric tends to minimize this error as defined in Eq. 2.7.

$$PSNR(SR, HR) = 10 \log_{10} \left(\frac{MAX^2}{MSE(SR, HR)} \right) \tag{2.7}$$

where MAX represents the highest possible value of a pixel in the image. Hence, if both images are similar in pixel-level then the PSNR will have a high value.

Even though higher PSNR values give the understanding of similar images, Santos et al. (2022) argued that this metric may damage face recognition systems. This happens because a high PSNR value leads to smoother faces, and the loss of facial marks leads to lower values of cosine similarity between images of the same face.

Finally, SSIM was proposed by Wang et al. (2004), this metric takes advantage of known characteristics of Human Visual Systems to compare local patterns of pixel intensities to evaluate luminance, contrast, and structure. In Eq. 2.8 it is possible to observe the general formula of the SSIM metric.

$$SSIM(SR, HR) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2\mu_y^2 + C_1)(\sigma_x^2\sigma_y^2 + C_2)} \tag{2.8}$$

Let x and y be images of size M x N, representing the SR and HR images in the case of this work. The function μ_x evaluates the luminance of an image and is defined in 2.9.

$$\mu_x = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} x(i,j) \tag{2.9}$$

The σ_x function evaluates the contrast of an image and is defined in 2.10.

$$\sigma_x = \left(\frac{1}{(N-1)(M-1)} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (x(i,j) - \mu_x)^2 \right)^{\frac{1}{2}} \tag{2.10}$$

The σ_{xy} function is used for structure comparison between two images and is defined in 2.11.

$$\sigma_{xy} = \frac{1}{(N-1)(M-1)} \sum_{i=0}^{M-1} \sum_{i=0}^{N-1} (x(i,j) - \mu_x)(y(i,j) - \mu_y) \tag{2.11}$$

Finally, C_1 and C_2 are included to avoid instability when the terms in the same denominator are close to zero, these terms are defined in a general form in Eq. 2.12.

$$C_x = (K_x L)^2 \tag{2.12}$$

where $K_x \ll 1$ is a small constant and L is defined as the maximum possible pixel value of an image.

Based on the above equations, it is possible to infer the need to maximize the SSIM metric in order to consider two images as similar. This is stated because if two images are similar then their means represented in Eq. 2.9 will tend to the same value while the variances presented in Eq. 2.10 and Eq. 2.11 will tend to lower values, thus maximizing the function.

2.4 CONCLUSION

This chapter introduced FSR as a domain-specific problem of ISR while classifying the existing methods developed to solve this task. It also covered the existing methods for artificially degrading an image once there are few datasets available with pairs of labeled LR and HR images with real-world degradation. Finally, it covered the three main metrics used to evaluate the results of current state-of-the-art FSR methods.

The next chapter will review the literature on FSR, presenting the state-of-the-art methods and using the classification presented in Section 2.1. It will also cover an overview of the datasets used by the community to build novel solutions.

3 RELATED WORKS

This chapter proposes to review existing methods on FSR, as well as the datasets used across different works. Section 3.1 presents a brief description of the datasets and a summary table, and Section 3.2 describes the SOTA approaches for FSR accompanied by tables summarizing the results obtained by each work.

3.1 DATASETS

The AR Face dataset proposed by Martinez and Benavente (1998) collected color images in 2 sessions separated by 14 days, in each session 13 pictures with 768x576 pixels were taken from 116 subjects. The collection was controlled in terms of equipment, distance, pose and illumination conditions, and no restrictions were applied on the wearing of the subjects.

Phillips et al. (1998) proposed the FERET dataset. This dataset collected gray images in 15 sessions, in each session 5-11 pictures with 256x384 pixels were taken from each subject with a total of 1199 subjects. The collection of the image was made in a semi-controlled environment since only the equipment was maintained the same across sessions. In this work, some individuals have a two-year lapse between each session, and not all subjects were present in more than one session.

On a different approach, Messer et al. (1999) collected color videos to build the XM2VTS dataset. The data was collected in 4 sessions separated by 5 months, in each session 2 shots were taken from each subject with a total of 295 subjects. The dataset was set up with control over the environment, but no control over the wearing of the subjects.

Roh and Lee (2007) also focused on collecting 1200 color videos but focused on acquiring color images from 1920 subjects with a resolution of 640x480 pixels. The proposed KFDB dataset had only 100 subjects present across the 3 sessions, each session was separated by one year. All the data were acquired in a controlled environment.

Instead of taking pictures in different sessions from multiple individuals, Huang et al. (2007) proposed to collect images from the internet. This approach resulted in the LFW dataset, composed of color images with 250x250 pixels from 5749 subjects. Since the images were obtained in the wild there was no control over the environment or any other restrictions.

Gao et al. (2008) also proposed a bigger dataset similar to the KFDB. Color images with a resolution of 360x480 pixels were taken from 1040 subjects, from which only 66 subjects did more than one session. The images were taken in a controlled environment with variations proposed by the authors.

On a classic approach, the FEI dataset was proposed by Thomaz and Giraldi (2010). Color images with 640x480 pixels were collected from 200 subjects. Similar to the previous work, the dataset was built in a controlled environment with proposed variations for each subject.

Upon the need for a dataset representing real-world scenarios, Grgic et al. (2011) proposed the SCface dataset. The dataset is composed of color images with multiple resolutions captured in an indoor uncontrolled environment by 5 surveillance cameras from 130 subjects.

Pinto et al. (2011) used a similar method to the one used to build the LFW dataset. Color images with multiple resolutions were collected in the wild from the internet forming the PubFig83 dataset containing 83 subjects.

Le et al. (2012) also collected color images in the wild from the internet, proposing the Helen dataset, which contains high-resolution images from multiple unidentified subjects. A different contribution in this dataset is the presence of feature location on each image.

Following the same approach, Yi et al. (2014) crawled over online content to create the CASIAWebFace dataset. This approach allowed the authors to obtain a large number of collected data, unfortunately, at the time of this research, the dataset was not found available online.

From a different perspective, Uzair et al. (2015) proposed to use of an indoor image system to collect hyperspectral images with 30x30 pixels in 4 sessions. The UWA dataset contains images from 70 subjects with 120 cubes each. The data were collected under a controlled environment with small variations in the pose.

While also collecting images from the internet, Liu et al. (2015) created the CelebA dataset containing color images taken in the wild with different resolutions from 10000 subjects. This dataset also contributes with the presence of 40 face attributes and 5 landmarks features for each image present in the dataset. Karras et al. (2017) then improved this dataset by proposing an algorithm to create high-quality images with 1024x1024 pixels from a subset of this dataset, creating the CelebA-HQ dataset.

Moschoglou et al. (2017) created the AgeDB dataset to support techniques where knowing the age of the subjects was important. Color images were collected online in the wild through crawling and selected only images where the identity, age, and gender of the subjects were well described. An average of 29 images were collected from each one of the 568 subjects.

In order to improve the CelebA-HQ dataset proposed previously, Karras et al. (2018) also collected color images online on the wild building the FFHQ dataset, a larger dataset with high-quality color images since each image has 1024x1024 pixels.

Cao et al. (2018) also took advantage of the computing power to crawl the internet to collect images with multiple resolutions in the wild. The authors built the largest dataset described in this work, the VGGFace2 dataset, containing an average of 362 color images from 9131 subjects.

On a different approach, Poster et al. (2021) proposed the collection of thermal images to build the ARL-VTF dataset. It was collect images from 395 subjects containing annotations such as identity, landmarks, and bounding boxes. The images were taken under a controlled environment with small variations in pose and wearing.

Table 3.1 summarizes the described datasets and the number of samples in each one. All the datasets previously described have the necessary information to be used for identification, except for Helen and FFHQ. All the datasets are also publicly available except by XM2VTS, KFDB, and CASIAWebFace.

Table 3.1: Common Datasets

Dataset	**Work**	**Samples**
AR Face	Martinez and Benavente (1998)	3016
FERET	Phillips et al. (1998)	14126
XM2VTS	Messer et al. (1999)	2360
KFDB	Roh and Lee (2007)	89380
LFW	Huang et al. (2007)	13233
CAS-PEAL-R1	Gao et al. (2008)	30900
FEI	Thomaz and Giraldi (2010)	2800
SCface	Grgic et al. (2011)	4160

Continued on next page

Table 3.1 – Continued from previous page

Dataset	Work	Samples
PubFig83	Pinto et al. (2011)	8300
Helen	Le et al. (2012)	2330
CASIAWebFace	Yi et al. (2014)	494414
UWA	Uzair et al. (2015)	4851
CelebA	Liu et al. (2015)	200K
CelebA-HQ	Karras et al. (2017)	30000
AgeDB	Moschoglou et al. (2017)	16488
FFHQ	Karras et al. (2018)	70K
VGGFace2	Cao et al. (2018)	3.31M
ARL-VTF	Poster et al. (2021)	500K

3.2 EXISTING METHODS

Baker and Kanade (2000) were among the first authors to propose treating facial reconstruction as its own class-based problem instead of making weak assumptions about the image to be reconstructed. The authors proposed an algorithm to learn the resolution enhancement function for lower levels of a Gaussian Pyramid, this was done by learning a prior on the derivatives of the high-resolution image and this information was then incorporated into the enhancement functions.

For generic images, Liu et al. (2001) proposed the use of helper images to learn how to reconstruct degraded images with statistical methods. The method used a two-step approach: first, it was obtained an optimal global face from the eigenspace by using the Gaussian Distribution to find similar results to the input image and by using a Markov Network to learn the statistical relationship between the global face image and local features, secondly, an optimal local feature image is inferred from the obtained global image by minimizing the energy of the Markov Network. Unfortunately, the authors did not present any quantitative results.

In a different approach, Gunturk et al. (2003) transferred the reconstruction task from pixel level to a lower dimensional face space, allowing the algorithm to construct the information directly in the low-dimensional domain reducing any unnecessary overhead. This was done by handling the observation noise and subspace representation error in the low-dimensional face subspace by using a Bayesian estimation of the feature vectors. Unfortunately, the authors did not present any global quantitative metric for the proposed method.

Wang and Tang (2005) introduced the idea that face reconstruction could be a transformation task between different image styles since previous works were based only on probabilistic methods. The authors used PCA to represent the structural similarity of face images in the feature space, since the method is applied to LR images it selected the right number of eigenvalues to extract the maximum amount of facial information.

Arguing that previous works did not well incorporate specific prior information about the images, Chakrabarti et al. (2007) proposed to extract this information using PCA by using a kernel to extract valuable information in a computationally efficient manner and incorporate this information within a *maximum a posteriori* framework. For this proposed approach, the authors also did not present a global quantitative metric to evaluate the results.

To Park and Lee (2008) typical interpolation approaches to generate SR images generally have poor results since these methods don't allow any new information to be included in the process. To address this problem the authors developed a novel approach consisting of two steps:

minimizing the error between the LR input image and a reference image by a linear combination of prototypes in LR images and applying the estimated coefficients to HR prototypes.

Liang et al. (2013) argued that the previous methods were mainly failing when modeling the global features of the image. To address this problem, the authors presented a method consisting of three steps: firstly the LR image is upsampled by interpolation, then it's applied Morphological Component Analysis to obtain the global approximation of the HR image and, finally, the detailed information to the estimated HR image is compensated by using the neighbor reconstruction of patches.

Innerhofer and Pock (2013) improved a previously built non-convex modeling method for image reconstruction by formulating a convex primal optimization problem and deriving a fast converging primal-dual algorithm with a globally optimal solution. The proposed method incorporates aligned face images prior to the reconstruction process by using the SiftFlow algorithm to densely align candidate images from a larger HR image dataset. The proposed method solves a generic saddle-point problem using a fast converging primal-dual algorithm.

Yang et al. (2013) proposed a face reconstruction algorithm that exploits domain-specific image structures since previous works focused on the generic super-resolution problem. The proposed algorithm uses landmark detection to locate facial components and contours, and process facial alignment in both frontal faces and those at different poses. To generate SR images from LR input images, the algorithm compares the aligned facial components of the input images with those of the training LR images and selects the LR exemplar images with the most similar components. The gradients of these exemplar images are preserved in reconstructing the output HR image. Unfortunately, the authors did not provide any global quantitative results for the proposed method.

Mei et al. (2020) used Cross-Scale Non-Local Attention to measure the correlation between low-resolution pixels and larger-scale patches in the LR image to improve the generated details on the facial region. This information is then used by each cell of a Recurrent Neural Network (RNN) to generate SR images.

Ying et al. (2021) also argued that the details generate by classical CNN methods are poor and proposed a novel approach able to predict the SR wavelet to obtain clearer facial images. The proposed approach uses a pre-trained semantic segmentation network to generate facial mask images during training, the creation of these masks is succeeded by the use of a linear low-rank convolution during the feature embedding and the use of skip connection in the wavelet coefficients prediction. Finally, a two-dimensional inverse discrete wavelet transform is utilized to reconstruct the HR image by using the predicted wavelet coefficients.

To solve this problem, Liu et al. (2021) used a Modal Regression-based Graph Representation to exploit the inherent topological structure for data representation, resulting in more accurate reconstruction coefficients. The authors also proposed the use of a modal regression-induced metric instead of using the classic least-square metric, since the classical approach is ineffective when dealing with the environment's noise.

Shi et al. (2022) proposed the use of regularized latent space exploration to address this problem. This approach uses a pre-trained GAN, during the iteration, SR faces are continually generated from a feasible latent space by the generator, and the generated images are evaluated in a way that the latent vector is gradually converged to the optimal solution.

In order to achieve better results, Chen et al. (2021a) based their work on the affirmation that some features are more important than others, hence the authors created an Attention-based Network able to adaptively bootstrap features related to the key facial structures. The authors created a Face Attention Unit which extends the original residual block by introducing a Spatial

Attention Branch, when this structure is stacked together it can continuously enhance important features in the SR image.

To focus on information-rich regions, Lu et al. (2021) used a Split Attention Network. It uses an Internal-feature Split Attention (ISA) mechanism to capture the information from these regions, then it uses an Internal-feature Split Attention Block (ISAB) to restore the facial texture details, next it is built an External-feature Attention Group (ESAG) based on two paths: one path cascades several ISABs to focus on facial texture details and the other path focuses on the facial structure information through ISA. Finally, a Split-attention In Split-attention Network is built on top of a cascade of several ESAGs for reconstructing photorealistic SR images.

Bao et al. (2022a) used a Cross-scale Dynamic Graph to exploit feature correlations allowing the network to treat different spatial regions with different attention. This was done by using a Channel Attention and Space Dynamic Graph that allows the model to focus on informative regions across different channels, this structure consists of two branches: a Channel Branch used to re-scale channel-wise features taking interdependencies into account and a Spatial Branch used to encode latent relationships between features patches.

Based on the idea that previous works tended to divide the LR images into grids, Liu et al. (2022) argued that this approach makes pixels that belong to the same shape mechanism be segmented into different patches. To address this issue it was proposed the Superpixel-guided Locality Quaternion Representation method, guarantees that local pixels belonging to the same structure are encoded simultaneously.

Observing that CNN approaches fail to capture global context on an image, (Liang et al., 2021) proposed the enhancement of the existing Swin Transformer Network, once Transformers have a better ability to capture the global-missing information. The proposed method is composed of three modules: the shallow feature extraction module which uses a Convolutional Layer and preservers low-frequency information, a deep feature extraction module responsible for local attention and learning cross-window interaction, and a reconstruction module which fuses the shallow and deep features for high-quality image reconstruction. Unfortunately, this work was only tested on non-facial datasets.

Wang et al. (2022d) used the same observation to work with face reconstruction through the use of Transformers. In this work, a shallow feature extractor is used to extract rich information that is then used to feed the global representation path and the local representation path. Both pieces of information are fused at the end of the pipeline using a Global-Local Aggregation Module.

Recently, Bao et al. (2023) proposed a novel Spatial Attention-guided CNN-Transform Aggregation Network. This novel approach allows the Transformer to interact with the entire process of feature extraction instead of using this information only at the reconstruction stage, allowing the Transformer to take advantage of the interaction between local and global information. Since Transformers tend to be cost-expensive in terms of computation, the authors also proposed the use of MLP during the upscaling phase to reduce the computational cost.

To address the cost problems, Zhang et al. (2022a) proposed an Efficient Long Range Attention Network tested on non-facial datasets. Initially, the network uses two successive shift convolutions to extract local features, then it uses a Group-wise Multi-scale Self-attention (GMSA) operator to construct different window sizes and calculates the self-attention separately, finally, it's proposed the use of a shared attention mechanism to accelerate the calculation for successive GMSA modules.

Immidisetti et al. (2021) approached the problem of reconstruction SR images from thermal LR images through the use of Axial-attention Blocks to model the global context while Convolutional Layers focused on the local features.

Following a similar technique, Jiang et al. (2022) explored the use of hyper-spectral images to generate SR images. Since there are few datasets available, the method is based on splitting the data into multiple groups to analyze each group using Spatial-Spectral Residual Blocks that are later aggregated in deeper layers of a CNN.

Zhang et al. (2022b) worked with video reconstruction proposing the focus on four key technical points: initial feature extraction, inter-frame information transfer, inter-frame alignment, and upsampling. Initially, it's designed as a U-shaped spanning feature extractor to lay a solid foundation when extracting initial features, then it's used a two-way transfer inter-frame information in order to fully utilize the time series and improve the quality of the SR image of each frame, finally, inter-frame alignment is performed in the form of residual learning for deformable convolution, and the accuracy of alignment is improved by using two residual learning with optical flow and bias reconditioning. Unfortunately, the tests were performed over non-facial datasets.

Since prior facial information is generally extracted from LR or SR images, (Ma et al., 2020) argued that this information is not fully exploited, and proposes the use of facial landmarks detection parallel to the reconstruction process. Using an RNN, the authors built two branches: one branch responsible for face recovery and the other branch responsible for landmark estimation in each recurrent step, making that previous outputs of each branch are fed into the other branch so that both branches can collaborate at every step of the process. Finally, the estimated landmark maps are used to generate multiple attention maps to integrate the landmark information into the network.

Following the intuition that prior information can be helpful when reconstructing facial regions, Li et al. (2021) used an Attribute Transformation Network to upsample the input LR image into multiple SR feature maps, next this information is incorporated into an intermediary SR face. Finally, it is used a Structure Enhancement Network which receives the intermediary SR image, that jointly extracts deep face features and estimates facial boundary heatmaps, fusing them to generate the final SR image by Spatial Attention Mechanism.

Wang et al. (2022a) proposed the use of prior information from the HR image to improve the quality of the network once it is fed with good prior information during the training phase. The method described in this work is composed of two networks: a Teacher Network that directly concatenates the LR images and the HR prior information and a Student Network that distillates prior knowledge from the Teacher Network making it able to exploit and capture facial prior.

Similar to previous works, Wang et al. (2021) also argued that applying shared kernels to different face regions is ineffective once it leads to the absence of important facial details. To address this problem the authors proposed a novel Heatmap-aware Convolution, consisting of three parts: a Common Feature Extraction Module that is divided into the Super-resolution Branch (SRB) and the Heatmap Estimation Branch (HEB). Both information from SRB and HEB are merged using a Heatmap-aware Block.

Following the intuition of using different weights depending on the facial regions, Wang et al. (2022c) proposed a novel Facial Region Classification Network. The proposed method first predicts all training samples using a single pre-trained network and then classifies the patches of all training samples into three categories based on the PSNR values: simple, medium, and hard. After obtaining the training samples of these three categories, the samples are used to train three reconstruction networks with different parameters to represent the corresponding patches. Finally, it is used a global classifier trained jointly with the three reconstruction networks to ensure that the patches of the input images in the inference phase are correctly classified by the proposed classifier and fed into the interconnected networks for reconstruction.

In order to exploit non-fixed size training patches to include the facial inherent structural properties and use the contextual information of the image, Gao et al. (2022) developed a Context-patch Representation Learning with Adaptive Neighbor Embedding (ANE). The authors use an ANE to select similar neighbors of the input contextual patch set in order to perform representation learning and attain the optimal representation weights. By the end of the process, the required SR patch can be then obtained by using the same representation weights over the relevant LR patches.

Cheng et al. (2021) argued that facial priors can not fully represent the image information and proposed the usage of identity information during the training process. The authors proposed the construction of a Dual-identity Dual-loop Network upon two closed loops: the first loop is responsible for generating SR images and the identity constraint plays the role of identity-preserving in the SR feature space, the second loop is used to learn the image degradation process to make sure that LR images preserve identity information.

Following a similar point of view Wang and Wong (2021) improved the existing PULSE method by using a pre-trained Cyclic Generative Adversarial Network (CGAN) to predict the difference between the reconstructed SR image and the GT identity. For this, the authors built two modules: a Predicted Difference Module responsible for training the CGAN to predict the difference of SR images generated by PULSE at each level of a Laplacian Pyramid and a Synthesis Module that uses the pre-trained CGAN to construct the difference of SR images. By the end of the process, it is performed the addition between the difference and the synthesized SR images in order to assist PULSE to achieve closer results to the real identity.

Arguing the classical networks lead to over-smoothed outputs and generate harmful artifacts to the face recognition task, Zhang and Ling (2021) proposed a Supervised Pixel-wise Generative Adversarial Network that used a supervised pixel-wise discriminator that focuses on the photorealism of each pixel on the generated SR image when comparing to GT image. The face recognition performance is then improved by incorporating a face identity prior to using both the LR image and its face features extracted from a pre-trained face recognition model.

Different from previous works, Bao et al. (2022b) performs the identity extraction via a Resolution-Robust Identity Knowledge Distillation Network, which consists of two streams: one stream encourages the distillation of resolution robust identity knowledge and the other uses the correlation information from the HR-HR stream to guide the learning in the LR-SR stream. This information is then used by a Texture and Identity Integration Network to incorporate it into the SR process.

Instead of using identity knowledge to guide the learning process, Li et al. (2020) proposed to use multiple reference HR images to do this operation. To select the set of HR reference images the author initially formulated the optimal guidance selection by assigning different weights to different face regions, next it was used the moving least square to alleviate the pose difference between images and the adaptive instance normalization to translate the illumination on the guidance images. Finally, the use of multiple Adaptive Spatial Feature Fusion Blocks to guide the reconstruction of the LR images in an adaptive and progressive manner.

Wang et al. (2022b) also proposed a multi-source and identity-agnostic framework to enhance the SR images. The developed approach consists of two parts: an encoder responsible for extracting local features followed by information mining that matches these features and fuse them into similar information in the feature space and a second encoder responsible for extracting global features followed by another information mining responsible for computing the self-similarity in the process of interacting local and global information across different scales.

Lu et al. (2022) used a similar approach, but instead of using multiple HR reference images, it uses a single reference image in a non-identity-agnostic approach. The HR image is used

to extract prior facial information and embed this information into the LR input images. Since this is identity-dependent the LR images are mapped into an LRMix image able to reconstruct SR images in the test phase in an identity-agnostic manner.

Since the manually downgraded datasets tend to have a gap between their images and the reference HR real-world images, (Chen et al., 2021b) used a re-expression framework to improve image homogenization. The authors proposed a novel three-expression framework, including a projection from LR to LR space, a projection from LR to HR space, and a projection from SR to HR space. The degradation gap between the real-world test LR image and the manually synthetic test LR image is minimized by the data re-expression in LR/LR space, and the degradation gap between the initially inferred SR image and the real-world HR image is minimized by the data re-expression in SR/HR space.

To improve the existing methods, Rajput (2022) used Average Filtering-based Data Fidelity and Locality Regularization in the objective function, allowing the framework to overcome the inadequacies of previously proposed SR methods that fail to handle heavier noises such as the Gaussian.

Even though the methods described in this section represent the SOTA techniques for FSR, only Lu et al. (2021), Chen et al. (2021a), Immidisetti et al. (2021), Ma et al. (2020), and Li et al. (2020) made their work public available for use.

Table 3.2: Statistical Methods

Dataset	**Work**	**Metric**	**Result**
FERET	Baker and Kanade (2000)	RMSE	0.05
XM2VTS	Wang and Tang (2005)	RMSE	0.06
	Park and Lee (2008)	MAE	0.04
	Park and Lee (2008)	SSIM	0.68
KFDB	Park and Lee (2008)	MAE	0.03
	Park and Lee (2008)	SSIM	0.80
CAS-PEAL-R1	Liang et al. (2013)	MSE	0.37
PubFig83	Innerhofer and Pock (2013)	PSNR	24.13
	Innerhofer and Pock (2013)	SSIM	0.75

Table 3.3: Deep Learning Methods

Approach	**Dataset**	**Degradation**	**Work**	**PSNR**	**SSIM**
General	CelebA	Bicubic	Lu et al. (2021)	27.59	0.7912
			Ying et al. (2021)	31.58	0.9494
			Shi et al. (2022)	23.52	0.5400
			Bao et al. (2022a)	27.40	0.7989
			Wang et al. (2022d)	27.87	0.7999
			Bao et al. (2023)	27.62	0.8041
	Helen	Bicubic	Chen et al. (2021a)	26.59	0.7716
			Bao et al. (2022a)	27.71	0.8306
			Bao et al. (2023)	27.87	0.8349
	FFHQ	Bicubic	Lu et al. (2021)	33.25	0.8738
			Wang et al. (2022d)	28.62	0.7972
	LFW	Bicubic	Ying et al. (2021)	33.34	0.9627
	ARL-VTF	Bicubic	Immidisetti et al. (2021)	16.57	0.5500

Continued on next page

Table 3.3 – Continued from previous page

Approach	Dataset	Degradation	Work	PSNR	SSIM
	CAS-PEAL-R1	Gaussian	Liu et al. (2021)	25.15	0.8060
	FEI	Gaussian	Liu et al. (2021)	27.40	0.8040
	AR Face	Gaussian	Liu et al. (2022)	24.53	0.7194
	UWA	Bicubic	Jiang et al. (2022)	35.23	0.8901
Prior-guided	CelebA	Bicubic	Ma et al. (2020)	27.37	0.7962
			Wang et al. (2021)	26.93	0.7759
			Li et al. (2021)	25.74	0.6779
			Wang et al. (2022a)	27.52	0.8057
	Helen	Bicubic	Ma et al. (2020)	26.69	0.7933
			Wang et al. (2022a)	26.79	0.7953
	FFHQ	Bicubic	Wang et al. (2022a)	26.39	0.7718
			Wang et al. (2022c)	28.39	0.7899
	LFW	Bicubic	Wang et al. (2022c)	27.23	0.7912
	FEI	Bicubic	Gao et al. (2022)	26.34	0.8284
			Lu et al. (2022)	35.98	0.9305
Identity-preserving	CelebA-HQ	Bicubic	Wang and Wong (2021)	32.25	0.9471
	Helen	Bicubic	Wang and Wong (2021)	35.17	0.9304
			Zhang and Ling (2021)	24.78	0.7360
			Bao et al. (2022b)	27.62	0.8255
	FEI	Bicubic	Cheng et al. (2021)	40.65	0.9650
	CelebA	Bicubic	Bao et al. (2022b)	27.18	0.7915
	LFW	Bicubic	Bao et al. (2022b)	29.24	0.8439
Reference	CelebA	Bicubic	Li et al. (2020)	26.39	0.9050
			Wang et al. (2022b)	30.06	0.8030
			Lu et al. (2022)	27.65	0.7946
	VGGFace2	Bicubic	Li et al. (2020)	24.34	0.8810
	CASIAWebFace	Bicubic	Li et al. (2020)	27.69	0.9210
	CAS-PEAL-R1	Gaussian	Chen et al. (2021b)	23.65	0.8906
	FFHQ	Bicubic	Lu et al. (2022)	28.49	0.8035
	FEI	Gaussian	Rajput (2022)	28.15	0.8268

3.3 CONCLUSION

This chapter covered a wide range of SOTA techniques for FSR with multiple approaches. The best results for the described works are highlighted in red in Tables 3.2, and 3.3. Even though the existing works achieved promising performance in terms of resolution metrics, most of them fail to evaluate the proposed algorithm against a identity identification scenario.

This work will evaluate the methods described in Chen et al. (2021a), Ma et al. (2020), and Li et al. (2020) to compare their identification abilities on real-world datasets. These methods were chosen based on the proposal of evaluating the implementation of works that are public available. The work proposed by Immidisetti et al. (2021) will not be included once the author's approach is used for thermal images and this scenario does not reflect the recognition problem in surveillance camera systems that will be evaluated by this work. Finally the work proposed by Lu et al. (2021) will also not be included due to time constraints to reproduce the experiments proposed by the authors.

4 PROPOSED APPROACH

This chapter discusses the proposed approach used to evaluate the SR networks used for an identification scenario. An illustration of the proposed approach is shown in Figure **??**.

It is important to notice that the identification evaluation is performed on a recognition scenario, i.e., each SR face reconstructed by the networks is compared to all the HR faces and, the most similar is considered as the same identity. A flowchart of the proposed evaluation pipeline can be found in Figure 4.1.

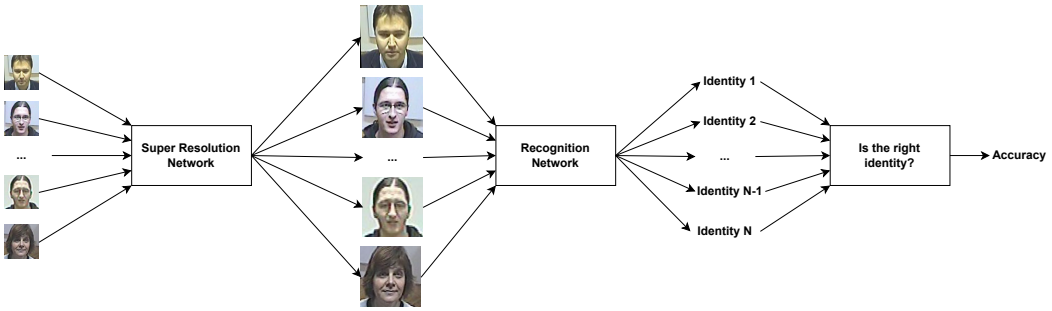


Figure 4.1: Proposed evaluation pipeline.

In Section 4.1, we briefly present the SR networks used and, in Section 4.2, we describe the used recognition networks. Finally, Section 4.3 presents the datasets used during the evaluation.

4.1 SUPER RESOLUTION NETWORKS

This section briefly describes each one of the SR networks proposed to be evaluated in this study.

4.1.1 DICNet and DICGAN

Ma et al. (2020) proposed a Deep Iterative Collaboration Network (DICNet) using a RNN architecture instead of classical deep generative models. The built network is based on the simultaneous work of two branches: one branch is responsible for face recovery, while the second learns how to properly estimate landmarks. An overview of the proposed architecture can be seen in Figure 4.2.

By using the proposed architecture, the authors argued that it is possible to better exploit the landmark's ability to help the face reconstruction process. This is based on the fact that at each training step, the network is fed with the combination of the landmarks and faces recovered in the previous step.

Beyond the recurrent architecture proposed, the authors also proposed an alternative approach to exploit prior facial knowledge. Instead of just concatenating the prior facial information with the recovered SR features, it was created an Attentive Fusion Module able to make better use of this information.

The module works by assuming that each landmark heatmap has a channel indicating the location of a specific landmark and groups these channels in one of the following: left eye, right eye, nose, mouth, or jawline. By fusing the learned landmarks it explicitly highlighted the

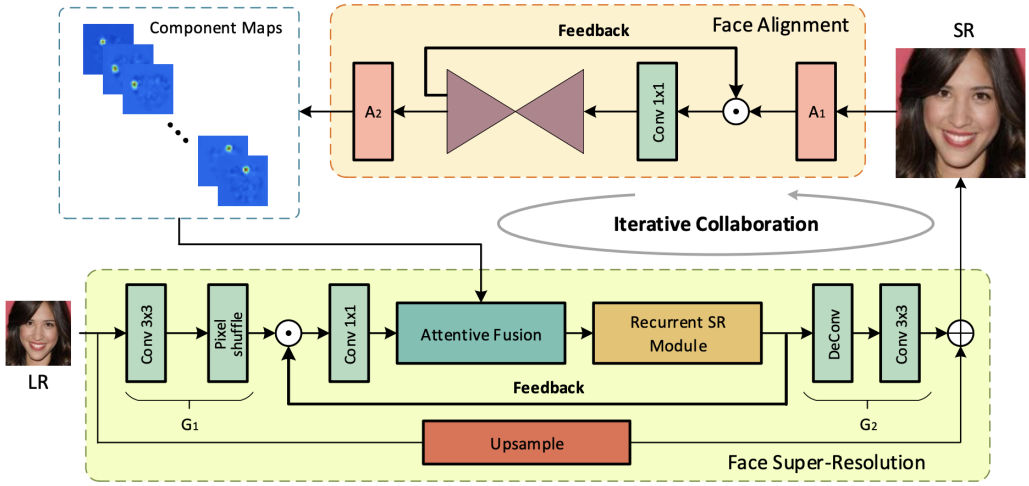


Figure 4.2: DICNet architecture. Extracted from Ma et al. (2020).

local structure of each facial part and the efficiency of the framework is improved by reducing the number of channels to learn. An overview of this process can be seen in Figure 4.3.

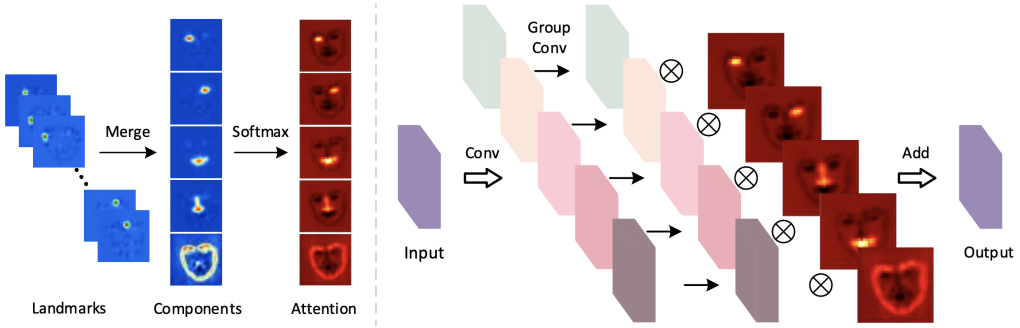


Figure 4.3: Landmark heatmap-extraction process. Extracted from Ma et al. (2020).

Finally, the network is trained by minimizing the pixel-level loss on the SR image and the alignment performed. Since the architecture of this network is based on a recurrent behavior, by combining the previous loss with the adversarial and perceptual losses the authors proposed a second network based on a GAN model called DICGAN.

Both networks were originally trained to receive as input a 16×16 image and output a 128×128 image (scale factor of 8). The networks were trained on the CelebA and Helen datasets proposed by Liu et al. (2015) and Le et al. (2012) respectively, using the bicubic degradation.

4.1.2 ASFFNet

Li et al. (2020) proposed an Adaptive Spatial Feature Fusion Network (ASFFNet) to enhance facial restoration by using a set of HR images as a reference to reconstruct the face of a given person. The network works by selecting from a set of HR images the most similar ones to the LR image to be reconstructed, the selected reference is then normalized to match the alignment and illumination of the LR image. Once this transformation is completed, the face restoration process is done with the proposed fusion blocks. An overview of this process can be seen in Figure 4.4.

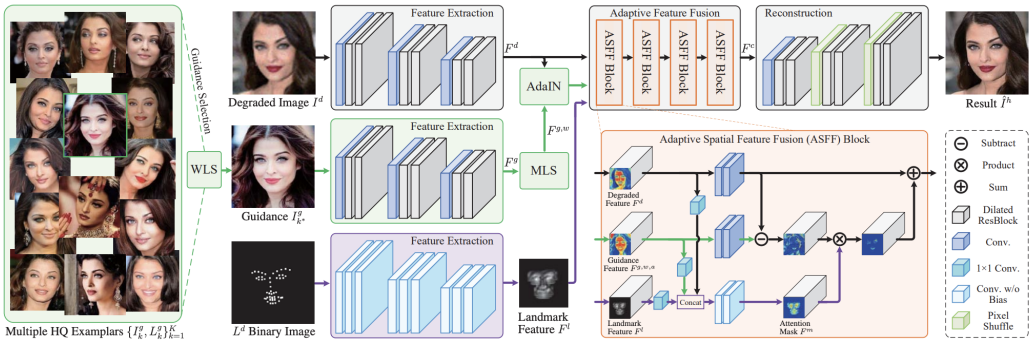


Figure 4.4: ASFFNet architecture. Extracted from Li et al. (2020).

The first step of this process is done by solving a Weighted Least Square (WLS) model that takes pose and expression into account to assign different weights to the given landmarks. Once these weights are assigned, the optimal HR reference image is obtained by finding the minimal affine distance. Once the guidance image is selected, it is possible to know that it matches the pose and expression of the LR image to be reconstructed, further it is necessary to match the alignment and illumination of both images.

The alignment correction is done by using the Moving Least Square (MLS) method to align both images in the feature space. Due to the differentiability of MLS it is also possible to train the feature extraction network to work collaboratively with this method. By treating the illumination problem as a style transfer problem, the authors also proposed to correct this feature by using the AdaIN network developed by Huang and Belongie (2017), adjusting the warped guidance feature to have similar illumination between both images.

In the last step, the warped guidance features are combined with the restored features to produce the SR output. Differently from other methods, the authors propose the use of multiple blocks instead of concatenating the features since this allows a progressive fusion of both information to output a better result.

Finally, the authors train the networks by minimizing two custom losses: the reconstruction loss and the photo-realistic loss. The first loss is given by the sum of the Mean Square Error and the Perceptual Loss, while the second loss is given by the sum of the Style Loss and the Adversarial Loss.

The network was originally trained to receive as input a 256×256 LQ image and outputs an image with its size based on the scale factor used (either x4 or x8). The network used in this study is a re-implementation of the original paper, trained to receive as input a 512×512 image and output an image with the same size. It was trained on a new dataset called CelebRef-HR proposed by the authors and made publicly available in Li et al. (2022b) using the ReDegNet degradation model proposed by Li et al. (2022a).

4.1.3 SPARNet and SPARNetHD

Chen et al. (2021b) proposed a Spatial Attention Residual Network (SPARNet) built upon novel Face Attention Units (FAUs). By integrating the FAUs into the residual blocks, the network can bootstrap features related to key facial structures (eyes, nose, mouth, etc.) and pay less attention to less feature-rich regions (shadows of the cheek, etc.). An overview of the proposed architecture can be seen in Figure 4.5.

The idea behind FAUs is that instead of hard-selecting attention maps to focus on specific regions, it is assigned a score between 0 a 1 for each spatial location on the map. This allows the network to learn the prediction of these maps by using gradient descent. This implies that

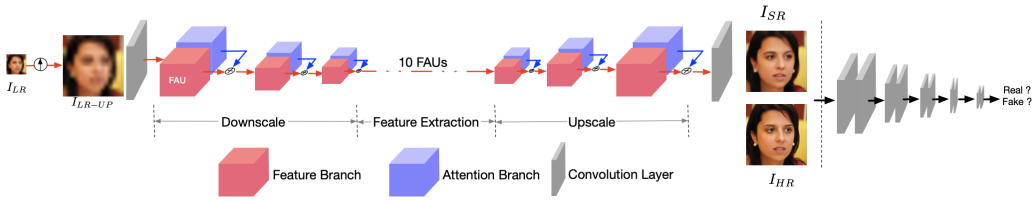


Figure 4.5: SPARNet architecture. Extracted from Chen et al. (2021a).

shallower layers focus on learning detailed structures such as hair, while deep layers can focus on learning coarse structures such as the mouth. The architecture of an FAU can be seen in Figure 4.6.

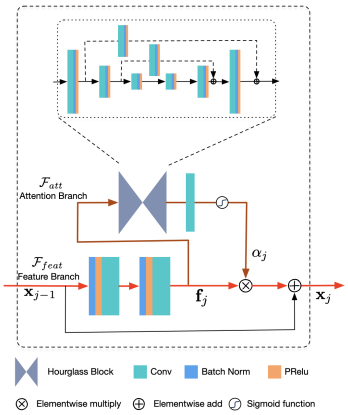


Figure 4.6: FAU architecture. Extracted from Chen et al. (2021a).

In each FAU, the Attention Branch is responsible for extracting multi-scale features and the Feature Branch focuses on learning the detailed features of the image. Finally, the network is trained by minimizing the pixel-level loss.

Expecting to reproduce real-world scenarios, Chen et al. (2021b) extended the proposed network to create SPARNetHD, based on the same network combined with a multi-scale discriminator. The author argued that by using the technique on multiple scales such as 512×512, 256×256, and 128×128 the quality of the SR images is improved.

The SPARNet network was initially trained to receive as input a 16×16 image and output an image with its size based on the scale factor used (e.g., a 32×32 input with a scale factor of 8 outputs a 256×256 image). The SPARNetHD receives a 512×512 input and outputs an image of the same size. Both networks were trained on the CelebA dataset proposed by Liu et al. (2015) using the bicubic degradation.

4.2 RECOGNITION NETWORK

Deng et al. (2019) proposed an Additive Angular Margin Loss (ArcFace) to be used while training CNNs for face recognition. The authors argued that by adding an additive angular margin penalty between the extracted features and the normalization's weights matrix it is possible to simultaneously enhance the intra-class compactness and inter-class discrepancy. An overview of the proposed approach can be seen in Figure 4.7.

The key idea behind adding the proposed penalty is to make the distance between classes in a hypersphere bigger, pushing similar images of the same class to be closer together. A

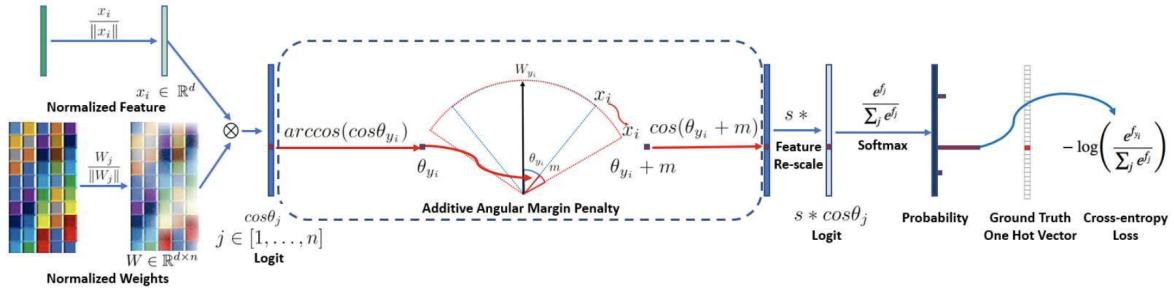
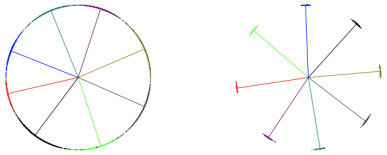


Figure 4.7: ArcFace architecture. Extracted from Deng et al. (2019).

geometric visualization of this process compared to the Softmax approach can be seen in Figure 4.8.



(a) Softmax (b) ArcFace

Figure 4.8: ArcFace geometric interpretation. Extracted from Deng et al. (2019).

The authors used the ResNet50 Network to generate the embeddings used to validate the novel loss. The network used to evaluate the proposed approach was trained using the CASIA (Yi et al., 2014), VGGFace2 (Cao et al., 2018), MS1MV2 [1], and DeepGlint-Face [2] datasets on a recognition scenario.

4.3 DATASETS

To be able to evaluate the proposed networks in real-world scenarios, on which LR images have variations in quality, pose, illumination, degradation, and others, the evaluation is performed on the SCFace and Quis-Campi datasets as inputs to the proposed pipeline. Section 4.3.1 covers how the SCFace dataset was created and its content, while Section 4.3.2 provides a similar description for the Quis-Campi dataset.

4.3.1 SCFace

The SCFace dataset was created by Grgic et al. (2011) to fulfill the needs of identity identification models in real-world scenarios. With that in mind, the database was built to be as close as possible to indoor scenarios using commercially available cameras with multiple qualities, variations in pose, illumination, and others.

The authors captured images from 130 subjects using 7 different cameras. While 5 of these cameras worked on visible light scenarios, 2 of them worked on an infrared range.

For each subject it was taken 3 photos from different pre-marked distances (1 meter, 2.6 meters, and 4.2 meters) with all the 7 cameras, totaling 21 images per subject in indoor

[1] No paper related to this dataset was referenced by the authors, claimed to be available at: http://trillionpairs.deepglint.com/overview

[2] No paper related to this dataset was referenced by the authors, claimed to be available at: http://trillionpairs.deepglint.com/overview

uncontrolled conditions of illumination. After that, each subject had 10 pictures taken with in a head rotation from -90º to 90º as a way to build the HR set to be used during identity identification, including the visible light mugshot. Finally, 1 picture from each subject was taken in a dark room to obtain a HR infrared mugshot as well.

In total, the dataset has 910 images from each camera in each one of the 3 distances mentioned, totaling 2730 LR images, from which 780 are in the infrared range. It also has 1300 HR images taken from different angles from each subject, 130 are mugshots, summed to 130 mugshots taken in infrared conditions. By the end, the dataset makes 4160 images available for testing. Some examples of these images can be seen in Figure 4.3.1.

Figure 4.9: SCFace dataset examples.

Since the proposed pipeline is be evaluated on a recognition scenario of visible lights. The initial SCFace dataset used during the tests removed all infrared images, as well as all the HR images but the mugshots taken. By the end, the used dataset has 1950 LR images and 130 HR mugshots to be used. It also kept two HR reference images from each subject facing the left and right sides to be used as enhancement references by the ASFFNet Network, totaling more 260 HR images.

4.3.2 Quis-Campi

The Quis-Campi dataset was created by Neves et al. (2017), with a similar purpose to the dataset described in 4.3.1. Going a step further, the proposed dataset in obtained in a totally unconstrained open environment, with multiple variations in pose, illumination, wearing, and others. All images are taken automatically at multiple distances up to 50 meters.

The author captures images from 268 subjects. For each subject, it was registered:

- 1 image of the background registration from where the subject would take the HR reference photo.

- 3 HR images of full body registration (frontal, left, and right sides).
- 6 HR videos of gait registration (from 6 different angles).
- 12 images, on average, taken in the open uncontrolled environment.
- 4 videos, on average, of the subject walking in the open uncontrolled environment.
- 4 videos, on average, of the background while the subject is walking outside.

In total, the dataset has 3548 LR images taken in multiple distances, poses, illumination, and others. It also has 268 HR images of the background registration for each subject and 804 HR full-body registrations. When talking about video captures it has 1608 HR videos of gait registration, as well as 1172 videos of the subject walking outside with matching 1172 videos of the background segmentation. Some examples of these images can be seen in Figure 4.10.

Figure 4.10: Quis-Campi dataset examples.

Since the proposed pipeline is evaluated on a recognition scenario of static images, all the videos made available by this dataset are initially discarded. For the HR reference only the frontal full-body registration image is kept and all the open-door LR images are kept as well. By the end, the used dataset has 3548 LR images and 268 HR frontal references to be used. It also kept two HR reference images from each subject facing the left and right sides to be used as enhancement references by the ASFFNet Network, totaling more 536 HR images.

Considering that the Quis-Campi dataset is in evolution and more images are added daily, as stated by the authors, it is important to notice that the numbers described previously might change after the publication of this work.

5 EXPERIMENTAL METHODOLOGY AND RESULTS

This chapter covers the methodology used to perform the proposed experiments, the scripts used for testing are available in GitHub. Section 5.1 covers the preprocessing techniques used on the proposed test datasets, Section 5.2 covers the common protocols used for identity identification, Section 5.3 presents the performed experiments and their results over SR images, and Section 5.4 discusses briefly the obtained results.

5.1 IMAGE PREPROCESSING

To make sure that all networks would be under the same scenario constraints, the first step developed during this work was to perform two common preprocessing techniques on all datasets: face detection and face alignment.

The goal of face detection involves three key steps: first, to eliminate images that may lack a visible face; second, to crop the face, ensuring that only the region of interest is retained for image reconstruction; and finally, to exclude any images containing more than one face, as this can interfere with identity identification evaluations. Examples of images removed at this stage can be seen in Figure 5.1.

Figure 5.1: Removed images. In (a), no visible faces and in (b), more than one face.

The face alignment necessity was mainly introduced by the fact that most of the SR networks are trained on aligned datasets and are not able to generalize to unaligned faces. Thus, it is necessary to pre-align the faces to obtain good results. Examples of cropped and aligned faces are visible in Figure 5.2.

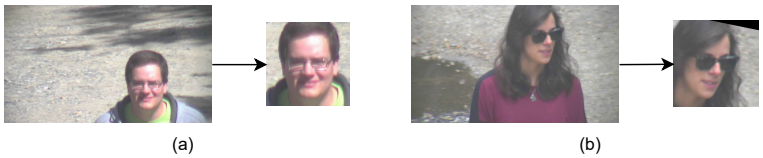


Figure 5.2: Preprocessed images. Image (a) had the face cropped and Image (b) had it cropped and slightly aligned.

To implement this technique it is used the approach proposed by Chen et al. (2021b), using the DLib's[1] face detection model and Scikit-image's[2] functions for alignment combined with the pre-trained weights provided by the authors. By default, the output images have a size of 512×512.

[1] http://dlib.net/
[2] https://scikit-image.org/

After the preprocessing, it is possible to observe that 570 LR images were removed from the SCFace dataset since it was not possible to identify faces on them due to the degradation. For this dataset, all the mugshots were kept.

The Quis-Campi dataset suffered the greatest impact in terms of quantity loss by having 3362 (94% of the LR images) removed after the preprocessing, this was caused by many factors such as: the high amount of images where the subjects were heading back to the camera, occlusion caused by pose or wearings such as glasses, and the presence of more than one subject in the image. Since the mugshots were also cropped from full frontal body registrations, differently from the previous dataset, the Quis-Campi had 5 HR references removed implying the removal of 5 subjects from the dataset.

5.2 IDENTIFICATION PROTOCOLS

Before proceeding to the evaluation scenarios elaboration it is necessary to establish the baseline test protocols. When working with identity identification multiple protocols can be used for evaluating the robustness of the proposed model. The most common are the open-set and closed-set protocols, adopted by this work.

To understand these protocols it is first necessary to understand the concept of probe and gallery sets. According to Du et al. (2022), the probe set refers to the images that need to be recognized through identification or verification, while the gallery set refers to the images registered in the recognition system with known identities. Considering that, the identification task refers to the network being able to define to which identity in the gallery set each image on the probe set belongs.

The open-set protocol states that some faces in the probe set are not necessarily in the gallery set, and this is generally the case when real-world applications are developed. The closed-set protocol then states the contrary, with the guarantee that each probe image has a corresponding identity in the gallery set.

In this work, each experiment is evaluated under the open-set and closed-set protocols. In this case, the SR images are part of the probe set, while the HR reference mugshots compose the gallery set.

5.3 EXPERIMENTS AND RESULTS

This section briefly discusses each one of the two proposed experiments and its results.

To fulfill the main goal of this work, understanding if the SR networks perform well in real-world scenarios, the pre-trained weights made available by the authors of each network were used for testing, without any additional fine-tuning.

Since some of the SR networks might not be able to successfully recover a face in the image, it is important to note that to perform a fair comparison, the images considered during the recognition step were only the images successfully recovered by all the networks evaluated in each experiment, i.e., if an image was successfully recovered by network A but not by network B then the image is discarded from the probe set. This constraint is further referenced as the full-set recoverability constraint.

Initially, Section 5.3.1 describes the statistical approach to validate the proposed experiments, Section 5.3.2 describes the first experiment by covering the results obtained by the SR networks that use a pre-defined scale factor to recover the faces on the LR images, and Section 5.3.3 describes the second experiment by presenting the results obtained by the networks that do not specify a scale factor and work by only enhancing the image quality.

5.3.1 Statistical Approach

A Paired Student's t-test was performed to fulfill the goal of this work with statistical relevance. The test is designed to compare the mean accuracy between the tests before and after applying an SR method to the LR images. The mean is calculated based on 30 samples obtained by randomly selecting a subset with 50% of a given dataset and performing the identification pipeline. The proposed hypothesis to be evaluated are as follows:

$$H_0 = \mu_m \leq 0$$
$$H_1 = \mu_m > 0$$

where μ_m represents the mean between the differences of the accuracies obtained before and after applying an SR method to the images. In other words, the null hypothesis states that using an SR method did not improve the accuracy at all or even harmed it, and the alternative hypothesis states that a statistically significant improvement was observed after applying an SR method.

Three main sets of experiments are performed: 32×32 inputs that are upsampled to 256×256 using SR, 64×64 inputs that are upsampled to 512×512 using SR, and 512×512 inputs that are enhanced without any upsampling. Considering these experiments, the SR methods are compared to two sets: images that have been upsampled using the nearest neighbors interpolation method, referred to as the Lower Bound (LB) set, and the images that are the direct output from the preprocessing step described in Section 5.1, referred as the Upper Bound (UB) set. Considering that the preprocessing step outputs 512×512 images, note that in the two first experiments, each image needs to be downsampled to match the SR's output size.

5.3.2 Networks with scale factors

In this experiment, only networks with a pre-defined scale factor are considered. As previously described, these networks take an LR input image and output an SR image depending on the scale factor, e.g., a 32×32 input under a scale factor of 8 outputs a 256×256 SR image.

In the literature, the most common scale factors are 2, 4, and 8. Unfortunately, the proposed SR networks only have the value 8 as a common scale factor accepted by the models.

This experiment is broken into two smaller tests: firstly, the networks are tested by receiving 32×32 inputs and, then they are tested by receiving 64×64 inputs. These inputs are obtained by resizing the LR images yielded by the preprocessing techniques described in Section 5.1.

Considering the full-set recoverability constraint, in this experiment, the SCFace dataset is composed of 1057 LR images under the open-set protocol and 1318 LR images under the closed-set protocol. The gallery set kept 130 HR references for the open-set protocol, while containing 104 HR references in the closed-set protocol.

Guided by the same constraint, the Quis-Campi dataset is composed of 186 LR images under the open-set and closed-set protocols. Meanwhile, the gallery set is formed by 263 HR references under the open-set protocol and only 182 HR references under the closed-set protocol.

The mean identification accuracy obtained for the first test, with 32×32 inputs, is shown in Table 5.1, where the best results are highlighted in red. Note that, even though SPARNet achieved the best results, all the methods had the same results when tested against the LB and UB sets.

Based on the hypothesis previously stated it is possible to observe that the null hypothesis is rejected in the LB comparison, supporting the idea that the accuracy is improved under this scenario. The same is not true for the UB comparison, where the null hypotheses could not be rejected.

Table 5.1: Accuracies obtained for 32×32 inputs at $\alpha = 0.05$ - Scale factor 8.

Protocol	Dataset	SR Method	Mean (%)	Std. Error	LB	UB
Open-Set	SCFace	Bicubic	37.99	2.36	H_1	H_0
		SPARNet	53.06	2.43	H_1	H_0
		DICNet	43.09	1.90	H_1	H_0
		DICGAN	39.89	2.34	H_1	H_0
	Quis-Campi	Bicubic	20.39	2.21	H_1	H_0
		SPARNet	33.47	1.42	H_1	H_0
		DICNet	32.54	2.66	H_1	H_0
		DICGAN	28.14	3.13	H_1	H_0
Closed-Set	SCFace	Bicubic	40.14	2.58	H_1	H_0
		SPARNet	53.12	2.46	H_1	H_0
		DICNet	43.60	2.18	H_1	H_0
		DICGAN	40.68	2.43	H_1	H_0
	Quis-Campi	Bicubic	20.50	2.69	H_1	H_0
		SPARNet	33.62	3.97	H_1	H_0
		DICNet	31.93	3.42	H_1	H_0
		DICGAN	28.93	3.63	H_1	H_0

The reason for the discrepancy between the lower and upper bounds can be seen in Figure 5.3. By manually resizing the image to a 32×32 input too much noise was added to the image, implying that any SR method can perform better than the LB baseline. Meanwhile, just resizing the preprocessed image to 256×256 (UB set) kept important facial details that can be used by the Recognition Network.

For the second test performed with 64×64 inputs, in Table 5.2 it is possible to observe an improvement in the mean accuracy, where the best results are also highlighted in red.

Note how, for the Quis-Campi dataset, SPARNet still holds the best results. In this case, a similar behavior to the one seen in the first test is observed: the noise added to the LB set is enough to make any SR method present a statistically significant improvement in accuracy, while the UB still holds the best facial characteristics to be used by the recognition network.

When analyzing the results for the SCFace dataset, a different conclusion is made: not only does using an SR method improve the accuracy in the UB comparison, but the classic bicubic interpolation method outperforms any network used. When inspecting the LB results, the same improvement holds for the bicubic interpolation approach but not for the other methods.

To understand this unexpected behavior, a closer look can be taken into the LB and UB accuracies: while for the open-set and closed-set protocols the LB set had 70.34% and 70.30% mean accuracy, respectively, the UB set had a mean accuracy of 68.49% and 69.35% for the same protocols, respectively. Based on this analysis, the data shown in Table 5.2 not only reflects the expected values but also leads to an important discovery: manually degrading the preprocessed image helped to remove the noise from it to the point where the accuracy when using an interpolation by the nearest neighbors method during upsample is still bigger than only using the original image.

Figure 5.4 presents some of the qualitative results that support the conclusions made for each dataset in the second test. It is particularly important to note how in the last column the LR (i.e., one of the LB images) seems to be less noisy than the preprocessed one (i.e., one of the UB images), reinforcing the assumption that the bicubic degradation followed by the reconstruction might help with noise reduction.

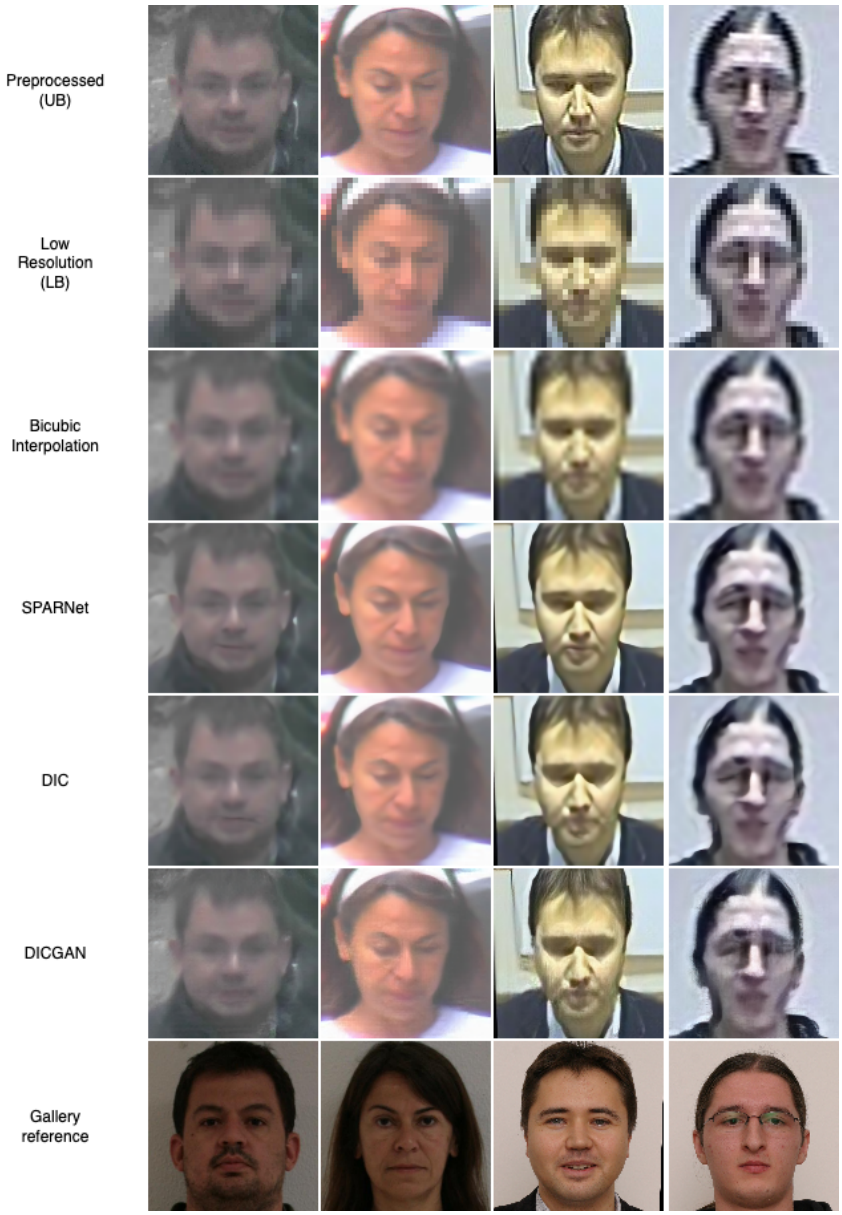


Figure 5.3: Examples of qualitative results having 32×32 input images and 256×256 pixels outputs. The first two and last two columns come from the Quis-Campi and SCFace datasets, respectively.

Table 5.2: Accuracies obtained for 64×64 inputs at $\alpha = 0.05$ - Scale factor 8.

Protocol	Dataset	SR Method	Mean (%)	Std. Error	LB	UB
Open-Set	SCFace	Bicubic	72.25	2.03	H_1	H_1
		SPARNet	69.25	2.36	H_0	H_1
		DICNet	69.10	2.36	H_0	H_1
		DICGAN	68.71	2.46	H_0	H_0
	Quis-Campi	Bicubic	43.23	3.05	H_1	H_0
		SPARNet	49.90	3.19	H_1	H_0
		DICNet	48.17	3.62	H_1	H_0

Continued on next page

Table 5.2 – Continued from previous page

Protocol	Dataset	SR Method	Mean (%)	Std. Error	LB	UB
		DICGAN	46.45	3.31	H_1	H_0
Closed-Set	SCFace	Bicubic	73.00	1.79	H_1	H_1
		SPARNet	69.98	2.19	H_0	H_1
		DICNet	69.86	1.93	H_0	H_1
		DICGAN	69.72	2.18	H_0	H_1
	Quis-Campi	Bicubic	43.91	4.39	H_1	H_0
		SPARNet	50.11	4.40	H_1	H_0
		DICNet	47.78	3.87	H_1	H_0
		DICGAN	46.70	4.17	H_1	H_0

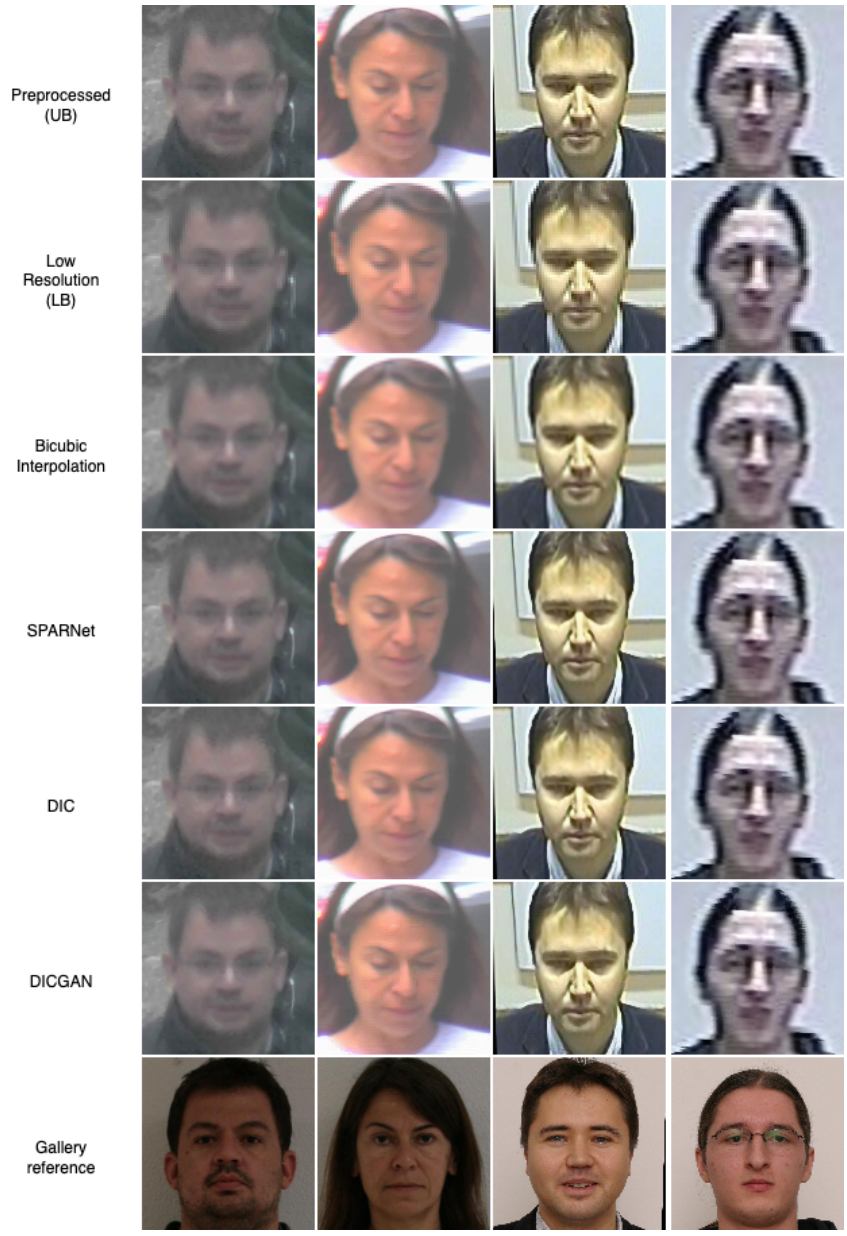


Figure 5.4: Examples of qualitative results having 64×64 input images and 512×512 pixels outputs. The first two and last two columns come from the Quis-Campi and SCFace datasets, respectively.

5.3.3 Enhancement networks

In this experiment, only the networks that do not take into account a scale factor for recovery are considered. Since the idea behind these networks is to only remove the noise from the input images and not upscale them, the LR 512×512 outputs produced by the preprocessing techniques of Section 5.1 are directly used as the probe set without the need of any further resizing. It is important to note that this set was not only formed by the restrictions defined in Section 5.1, but also included the ability to recognize the faces when subjects were facing sideways on the HQ reference images, causing a drop in the number of individuals available for identification to 114 on SCFace and 239 on the Quis-Campi.

Considering the full-set recoverability constraint, in this experiment, the SCFace dataset is composed of 763 LR images under the open-set and closed-set protocols. Meanwhile, the gallery set is formed by 130 HR references under the open-set protocol and 102 HR references under the closed-set protocol.

Based on the same constraint, the Quis-Campi dataset is composed of 131 LR images under the open-set and closed-set protocols. For the gallery set, the dataset is formed by 263 HR references under the open-set protocol and 130 HR references under the closed-set protocol.

The results for this experiment can be seen in Table 5.3, where the best results are highlighted in red. Notice that, similar to the described in Section 5.3.2, no method can outperform the upper bound in the Quis-Campi dataset even with SPARNetHD holding high accuracy values.

For the SCFace dataset, it is possible to observe that the methods presented improvements in the mean accuracy when the UB comparison is made, thus rejecting the null hypothesis.

Table 5.3: Accuracies obtained for 512×512 inputs at $\alpha = 0.05$.

Protocol	**Dataset**	**SR Method**	**Mean (%)**	**Std. Error**	**Upper Bound**
Open-Set	SCFace	SPARNetHD	75.66	2.10	H_1
		ASFFNet	63.91	2.46	H_0
	Quis-Campi	SPARNetHD	60.05	4.46	H_0
		ASFFNet	32.42	3.28	H_0
Closed-Set	SCFace	SPARNetHD	76.34	2.79	H_1
		ASFFNet	64.32	3.04	H_0
	Quis-Campi	SPARNetHD	61.01	4.10	H_0
		ASFFNet	33.49	4.97	H_0

Considering that for this set of experiments, no image needed to be resized to then be upscaled, there is no LB set to compare as in the previous Section.

Qualitative results for the inputs described for this experiment can be seen in Figure 5.5. In this case, it is possible to observe that the facial structures are better defined when compared to the SR images described in Section 5.3.2 even though some deformations are present.

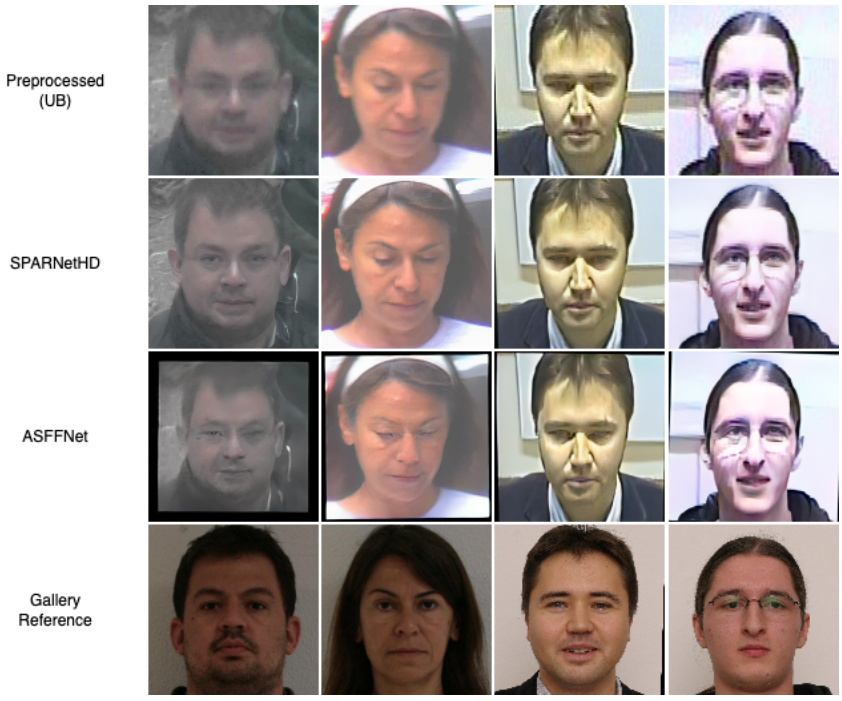


Figure 5.5: Examples of qualitative results for 512×512 inputs. The first two and last two columns come from the Quis-Campi and SCFace datasets, respectively.

5.4 DISCUSSION

This Chapter discussed the protocols employed during the evaluation of the proposed experiments, as well as the statistical approach to the problem. It was shown that manually degrading images to too small sizes to fit a dataset to be upsampled, harms the final accuracy obtained over a dataset. In counterpart, when this process was applied to just the right size the degradation can help to remove the noise from the original input, improving the accuracy obtained. Furthermore, using classical bicubic interpolation for upsampling and image enhancement methods showed better results than the upsampling SR approaches.

6 CONCLUSION

The SOTA methods developed over the past years for the SR problem have been widely discussed in this work. Due to the lack of work shareability in this field, the most promising SOTA methods could not be reproduced by this work, and less powerful networks had to be evaluated. By using the identification accuracy as the leading metric to evaluate the proposed papers on real-world scenarios, it was demonstrated that these networks can be helpful in the identification problem in highly degraded datasets, but the same does not hold for real-world datasets that have better equipment and higher resolution images. Furthermore, it was exemplified that the classical approach of bicubic interpolation for upsampling and the use of enhancement networks tend to output better results than the evaluated SR methods.

It is encouraged as the first step of future works to investigate deeper how using a manual bicubic degradation approach, followed by the nearest neighbors interpolation upsampling, affects other SR pipelines given the fact that it helped to remove noise from the highly noisy datasets.

As a second step, it is recommended to explore different recognition networks considering that, for LR images, the network architecture can have a significant impact on the final accuracy. In the same direction, it is suggested to test different face detection methods during the pre-processing step to avoid removing too many LR images such as what happened to the Quis-Campi dataset. As an alternative solution to this problem, it is possible to engage in manual face detection for each image in the dataset during this step.

REFERENCES

Baker, S. and Kanade, T. (2000). Hallucinating faces. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 83–88.

Bao, Q., Gang, B., Yang, W., Zhou, J., and Liao, Q. (2022a). Attention-driven graph neural network for deep face super-resolution. *IEEE Transactions on Image Processing*, 31:6455–6470.

Bao, Q., Liu, Y., Gang, B., Yang, W., and Liao, Q. (2023). Sctanet: A spatial attention-guided cnn-transformer aggregation network for deep face image super-resolution. *IEEE Transactions on Multimedia*, pages 1–12.

Bao, Q., Zhu, R., Gang, B., Zhao, P., Yang, W., and Liao, Q. (2022b). Distilling resolution-robust identity knowledge for texture-enhanced face hallucination. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 6727–6736, New York, NY, USA. Association for Computing Machinery.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*.

Chakrabarti, A., Rajagopalan, A. N., and Chellappa, R. (2007). Super-resolution of face images using kernel pca-based prior. *IEEE Transactions on Multimedia*, 9(4):888–892.

Chen, C., Gong, D., Wang, H., Li, Z., and Wong, K.-Y. K. (2021a). Learning spatial attention for face super-resolution. *IEEE Transactions on Image Processing*, 30:1219–1231.

Chen, L., Pan, J., Jiang, J., Zhang, J., Han, Z., and Bao, L. (2021b). Multi-stage degradation homogenization for super-resolution of face images with extreme degradations. *IEEE Transactions on Image Processing*, 30:5600–5612.

Cheng, F., Lu, T., Wang, Y., and Zhang, Y. (2021). Face super-resolution through dual-identity constraint. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Deng, J., Guo, J., Niannan, X., and Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *CVPR*.

Du, H., Shi, H., Zeng, D., Zhang, X.-P., and Mei, T. (2022). The elements of end-to-end deep face recognition: A survey of recent advances. *ACM Comput. Surv.*, 54(10s).

Gao, G., Yu, Y., Lu, H., Yang, J., and Yue, D. (2022). Context-patch representation learning with adaptive neighbor embedding for robust face image super-resolution. *IEEE Transactions on Multimedia*, pages 1–11.

Gao, W., Cao, B., Shan, S., Chen, X., Zhou, D., Zhang, X., and Zhao, D. (2008). The cas-peal large-scale chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 38(1):149–161.

Grgic, M., Delac, K., and Grgic, S. (2011). Scface — surveillance cameras face database. *Multimedia Tools and Applications*, 51(3):863–879.

Gunturk, B., Batur, A., Altunbasak, Y., Hayes, M., and Mersereau, R. (2003). Eigenface-domain super-resolution for face recognition. *IEEE Transactions on Image Processing*, 12(5):597–606.

Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst.

Huang, X. and Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization.

Immidisetti, R., Hu, S., and Patel, V. M. (2021). Simultaneous face hallucination and translation for thermal to visible face verification using axial-gan. In *2021 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–8.

Innerhofer, P. and Pock, T. (2013). A convex approach for image hallucination.

Jiang, J., Wang, C., Liu, X., Jiang, K., and Ma, J. (2022). From less to more: Spectral splitting and aggregation network for hyperspectral face super-resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 266–275.

Jiang, J., Wang, C., Liu, X., and Ma, J. (2021). Deep learning-based face super-resolution: A survey. *ACM Comput. Surv.*, 55(1).

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196.

Karras, T., Laine, S., and Aila, T. (2018). A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948.

Keys, R. (1981). Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(6):1153–1160.

Le, V., Brandt, J., Lin, Z., Bourdev, L., and Huang, T. S. (2012). Interactive facial feature localization. In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., and Schmid, C., editors, *Computer Vision – ECCV 2012*, pages 679–692, Berlin, Heidelberg. Springer Berlin Heidelberg.

Li, M., Zhang, Z., Yu, J., and Chen, C. W. (2021). Learning face image super-resolution through facial semantic attribute transformation and self-attentive structure enhancement. *IEEE Transactions on Multimedia*, 23:468–483.

Li, X., Chen, C., Lin, X., Zuo, W., and Zhang, L. (2022a). From face to natural image: Learning real degradation for blind image super-resolution. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, page 376–392, Berlin, Heidelberg. Springer-Verlag.

Li, X., Li, W., Ren, D., Zhang, H., Wang, M., and Zuo, W. (2020). Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2703–2712.

Li, X., Zhang, S., Zhou, S., Zhang, L., and Zuo, W. (2022b). Learning dual memory dictionaries for blind face restoration.

Li, Y., Chen, H., Li, T., and Liu, B. (2023). Ddnsr: a dual-input degradation network for real-world super-resolution. *Pattern Analysis and Applications*.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., and Timofte, R. (2021). Swinir: Image restoration using swin transformer. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1833–1844.

Liang, Y., Xie, X., and Lai, J.-H. (2013). Face hallucination based on morphological component analysis. *Signal Processing*, 93(2):445–458.

Liu, C., Shum, H.-Y., and Zhang, C.-S. (2001). A two-step approach to hallucinating faces: global parametric model and local nonparametric model. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I.

Liu, L., Chen, C. L. P., and Wang, Y. (2021). Modal regression-based graph representation for noise robust face hallucination. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13.

Liu, L., Tang, X., Chen, C. P., Cai, L., and Lan, R. (2022). Superpixel-guided locality quaternion representation for color face hallucination. *Information Sciences*, 609:565–577.

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

Lu, T., Wang, Y., Zhang, Y., Jiang, J., Wang, Z., and Xiong, Z. (2022). Rethinking prior-guided face super-resolution: A new paradigm with facial component prior. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15.

Lu, T., Wang, Y., Zhang, Y., Wang, Y., Wei, L., Wang, Z., and Jiang, J. (2021). Face hallucination via split-attention in split-attention network. In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 5501–5509, New York, NY, USA. Association for Computing Machinery.

Ma, C., Jiang, Z., Rao, Y., Lu, J., and Zhou, J. (2020). Deep face super-resolution with iterative collaboration between attentive recovery and landmark estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5568–5577.

Martinez, A. and Benavente, R. (1998). *The AR Face Database: CVC Technical Report, 24*. CVC.

Mei, Y., Fan, Y., Zhou, Y., Huang, L., Huang, T. S., and Shi, H. (2020). Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5689–5698.

Messer, K., Matas, J., Kittler, J., Luettin, J., and Maître, G. (1999). Xm2vtsdb: The extended m2vts database. In *Physics Department*.

Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., and Zafeiriou, S. (2017). Agedb: the first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop*, page 5.

Neves, J. C., Moreno, J. C., and Proença, H. (2017). Quis-campi: an annotated multi-biometrics data feed from surveillance scenarios. *IET Biometrics*, 7:371–379.

Park, J.-S. and Lee, S.-W. (2008). An example-based face hallucination method for single-frame, low-resolution facial images. *IEEE Transactions on Image Processing*, 17(10):1806–1816.

Phillips, P., Wechsler, H., Huang, J., and Rauss, P. J. (1998). The feret database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing*, 16(5):295–306.

Pinto, N., Stone, Z., Zickler, T., and Cox, D. (2011). Scaling up biologically-inspired computer vision: A case study in unconstrained face recognition on facebook. In *CVPR 2011 WORKSHOPS*, pages 35–42.

Poster, D., Thielke, M., Nguyen, R., Rajaraman, S., Di, X., Fondje, C. N., Patel, V. M., Short, N. J., Riggan, B. S., Nasrabadi, N. M., and Hu, S. (2021). A large-scale, time-synchronized visible and thermal face dataset.

Rajput, S. S. (2022). Gaussian noise robust face hallucination via average filtering based data fidelity and locality regularization. *Applied Intelligence*.

Roh, M.-C. and Lee, S.-W. (2007). Performance analysis of face recognition algorithms on korean face database. *International Journal of Pattern Recognition and Artificial Intelligence*, 21(06):1017–1033.

Santos, M. D., Laroca, R., Ribeiro, R. O., Neves, J., Proença, H., and Menotti, D. (2022). Face super-resolution using stochastic differential equations. In *2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, volume 1, pages 216–221.

Shi, R., Zhang, J., Li, Y., and Ge, S. (2022). Regularized latent space exploration for discriminative face super-resolution. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2534–2538.

Thomaz, C. E. and Giraldi, G. A. (2010). A new ranking method for principal components analysis and its application to face image analysis. *Image and Vision Computing*, 28(6):902–913.

Uzair, M., Mahmood, A., and Mian, A. (2015). Hyperspectral face recognition with spatiospectral information fusion and pls regression. *IEEE Transactions on Image Processing*, 24(3):1127–1137.

Wang, C., Jiang, J., and Liu, X. (2021). Heatmap-aware pyramid face hallucination. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6.

Wang, C., Jiang, J., Zhong, Z., and Liu, X. (2022a). Propagating facial prior knowledge for multitask learning in face super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7317–7331.

Wang, R., Jian, M., Smith, P., and Yu, H. (2022b). Face super-resolution based on multi-source references. In *2022 15th International Conference on Human System Interaction (HSI)*, pages 1–5.

Wang, W. and Wong, H.-C. (2021). Multi-level difference repair architecture for face hallucination. *IEEE Signal Processing Letters*, 28:2048–2052.

Wang, X. and Tang, X. (2005). Hallucinating face by eigentransformation. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35(3):425–434.

Wang, Y., Lu, T., Wang, Y., and Wang, Z. (2022c). Classifying facial regions for face hallucination. *IEEE Signal Processing Letters*, 29:2392–2396.

Wang, Y., Lu, T., Zhang, Y., Wang, Z., Jiang, J., and Xiong, Z. (2022d). Faceformer: Aggregating global and local representation for face hallucination. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Wu, C.-H. and Chang, H.-H. (2012). Gaussian noise estimation with superpixel classification in digital images. In *2012 5th International Congress on Image and Signal Processing*, pages 373–377.

Yang, C.-Y., Liu, S., and Yang, M.-H. (2013). Structured face hallucination. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1099–1106.

Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014). Learning face representation from scratch.

Ying, L., Dinghua, S., Fuping, W., Pang, L. K., Kiang, C. T., and Yi, L. (2021). Learning wavelet coefficients for face super-resolution. *The Visual computer*, 37(7):1613–1622.

Yue, L., Shen, H., Li, J., Yuan, Q., Zhang, H., and Zhang, L. (2016). Image super-resolution: The techniques, applications, and future. *Signal Processing*, 128:389–408.

Zhang, M. and Ling, Q. (2021). Supervised pixel-wise gan for face super-resolution. *IEEE Transactions on Multimedia*, 23:1938–1950.

Zhang, X., Zeng, H., Guo, S., and Zhang, L. (2022a). Efficient long-range attention network for image super-resolution. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, page 649–667, Berlin, Heidelberg. Springer-Verlag.

Zhang, Z., Guo, H., Ren, S., and Guo, K. (2022b). Learning inter-frame information for space-time video super-resolution. In *2022 Tenth International Conference on Advanced Cloud and Big Data (CBD)*, pages 139–144.