Deep Learning Based Image Super Resolution

by

Xiang Wang
Lakehead University

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Computer Science

Lakehead University

Deep Learning Based Image Super Resolution

by

Xiang Wang
Lakehead University

Supervisory Committee

Dr. Shan Du, Supervisor
(Department of Computer Science, Lakehead University, Canada)

Dr. Yimin Yang, Co-Supervisor
(Department of Computer Science, Lakehead University, Canada)

Dr. Jinan Fiaidhi, Departmental Member
(Department of Computer Science, Lakehead University, Canada)

Dr. Wilson Wang, External Member
(Department of Mechanical Engineering, Lakehead University, Canada)

ABSTRACT

Image super resolution is one of the most significant computer vision researches aiming to reconstruct high resolution images with realistic details from low resolution images. In the past years, a number of traditional methods intended to produce high resolution images. Recently, Deep Convolutional Neural Networks (DCNNs) have developed rapidly and achieved impressive progress in the computer vision area. Benefiting from DCNNs, the performance of image super resolution has improved compared with traditional methods. However, there still exists a large gap between the results of current methods and the real-world high resolution quality.

In this thesis, we leverage the techniques of DCNNs to develop image super res olution models for generating satisfactory high resolution images. There are several proposed methods in this thesis to satisfy different super resolution scenarios. Our proposed methods are based on Generative Adversarial Networks (GANs), leading to powerful generative ability and effective discriminative learning. To breakthrough current bottlenecks, we design novel architectures for generator and discriminator, and involve new optimization strategies to improve the learning stability of the mod els. In order to improve the generalization ability of proposed methods, we conduct two mainstream super resolution tasks, namely face image hallucination and natu ral image super resolution. All the proposed components of our methods result in promising super resolution performance for these tasks.

Not only handling the supervised super resolution task, we also investigate the more challenging problem, namely the unsupervised image super resolution task where the paired high resolution image and low resolution image data are unavailable. To evaluate the performance of our methods in different scenarios, we conduct exten sive experiments on several benchmark datasets to study each method separately. Compared to state-of-the-art methods, our methods are able to achieve superior per formance both quantitatively and qualitatively.

ACKNOWLEDGEMENTS

I would like to express my sincerest gratitude to the following people:

First of all. My supervisor, Dr. Shan Du, for her persistent guidance and uncon

ditional support throughout my Master study. She has supervised me with endless inspiration, immense knowledge, and great patience, which I will always appreciate. She offered me the valuable opportunity to study on computer vision and machine learning research and provided me with constructive advice no matter what diffi culties I encountered. I especially appreciate the freedom and trust she gave me to explore my own research topic in the master journey. She was not only an excellent supervisor but also a deep friend to me, and I was greatly honourable to study under her supervision and forever thankful for all I have learned from her.

My co-supervisor, Dr. Yimin Yang, for his invaluable suggestion and constant encouragement in my Master period. His rigorous and enthusiastic scientific spirit have deeply motivated and inspired me to overcome challenges during my thesis study. His professional techniques and profound advices guided me to complete this thesis. I learned a lot from him especially in the deep learning area. He taught me how to conduct academic research and how to solve thorny problems. I was very lucky to be his student and cooperate with him.

Last but not least, I would like to deeply thank my family. They have always given me endless patience and powerful support. Their love and appreciation always encourage me to go further.

“The feature belongs to those who believe in the beauty of their dreams.” -

Eleanor Roosevelt

PUBLICATIONS

Xiang Wang, Shan Du, Yimin Yang, Qixiang Pang, Qiang Tang. End-to-End Gen erative Adversarial Face Hallucination Through Residual In Internal Dense Network, under review of European Signal Processing Conference, 2021.

Xiang Wang, Shan Du, Yimin Yang. Semantic Encoder Guided Generative Ad

versarial Face Ultra-Resolution Network, in preparation (towards IEEE Trans. on Industrial Informatics), 2021.

Xiang Wang, Shan Du, Yimin Yang. Real-World Image Super Resolution via Unsu pervised Bi-directional Cycle Domain transfer Learning based Generative Adversarial Network, in preparation (towards Neurocomputing), 2021.

Terence Chow *, Xiang Wang *, Yimin Yang, Shan Du. Decision Fusion-based Ensemble Deep Network For Hybrid Scene Recognition, under review of IEEE Inter national Conference on Systems, Man and Cybernetics, 2021.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

Over the past decades, the demands for High Resolution (HR) images have increased dramatically. Image resolution indicates the details contained in an image. As the digital images comprise elements of pixels, it is common to use the density of pixels in per unit image area to indicate the spatial resolution. Thus, the images with high resolution represent that pixel densities of images are high and more image details can be shown. In the modern digital world, HR images are widely applied to many image applications, such as visual surveillance [58, 62], object classification [15, 104], medical diagnosis [35, 36] and remote sensing [28, 33], since HR images usually bring pleasant visual effect and provide accurate details for image analysis. As shown in Figure 1.1, we give some sample images from an HR image to its LR images with different downsampling factors. It clearly shows that the resolution gradually decreases from image in column (a) to image in column (d) while the image details dramatically

deteriorate by degrees. Specifically, compared to ground truth (column (a) in Figure 1.1), it is difficult to visualize the textures and contents of the LR image in column (d). Thus, it is necessary to apply HR image with clear details to real world applications, achieving better performance. However, the HR images usually cannot be obtained in many scenes because of poor image acquisition devices and unsuitable environment, such as low quality cameras and gloomy environment. Therefore, the acquired images taken by such equipment and scenes normally are low resolution with sensor noises and unexpected artifacts.

Figure 1.1: Some sample images of an HR image and its LR versions with different downsampling factors. Column (a) is the ground truth (256×256 pixels). Column (b) is an LR image downsampled by factor 2× (128×128 pixels). Column (c) shows the LR image with downsampling factor 4× (64×64 pixels). And column (d) presents the LR image with downsampling factor of 8× (32×32 pixels).

To address the problem of poor image quality, many solutions have been pro posed. There are two normal ways to acquire HR images: the first solution is to use sophisticated hardware equipment and the other one relies on image enhancement algorithms. From the hardware aspect, many sensor device manufacturers make a great effort to update digital sensors and optical components to capture clear infor mation of images. Nevertheless, there still exist unsolvable problems even though we have powerful devices. For example, the inevitable factors, such as undeveloped tech niques and unacceptable costs, severely hinder the development of hardware which can obtain HR images. In addition, it is impracticable to update all hardware com ponents in the existing imaging systems and devices. Under these conditions, many researchers consider the second solution, image enhancement algorithms. The second solution aims to design reliable algorithms to recover missing contents of HR image from LR image. Image enhancement algorithms are super feasible solutions in the real world. In this case, the algorithms which reconstruct the HR image from its LR

version are named super-resolution algorithms. Because of promising reconstruction performance and acceptable cost, the super-resolution algorithms have received wide attention and became one of the most popular research directions in the image pro cessing area. In this thesis, we mainly investigate several effective methods from the super-resolution algorithms perspective.

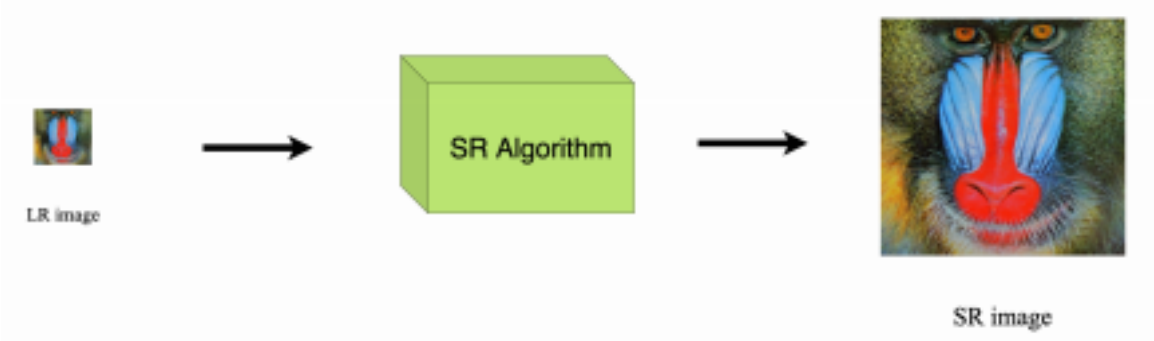

Figure 1.2: An example of the super-resolution mechanism. The left image is a low resolution image at the size of 64 × 64; the right image is a SR image generated by the SR algorithm at the size of 256 × 256.

Single Image Super Resolution (SISR) aims to recover from a given LR image with blurry textures to the high quality HR image with sharp details. In Figure 1.2, the basic SISR mechanism is demonstrated. The left is the LR image which can be regarded as the downscaled version of its HR image and the right is the final SR result. The SR algorithm in Figure 1.2 can be seen as the SR black box, which receives and further upscales the small sized LR image to the final HR image. There are many SR methods can be put in this SR block box, such as SRCNN [17], RDN [98], SRDenseNet [72]. In the last decades, various SR methods have been proposed. They can be divided into two main categories: traditional methods and deep learning-based methods. The details of these methods can refer to chapter 2.

1.2 Motivation

In this thesis, the motivation of our studies is the demand of image super-resolution tasks in computer vision applications. As mentioned in above sections, the HR im ages with high pixel density are in great demand in the real world applications, which can provide high resolution images to achieve better performance in these applica tions. Thus, the study of image super-resolution problems is essential. A large effort

has been made in our work to investigate remarkable SR methods to handle SISR problems.

1.3 Problem Challenges

As mentioned above, single image super-resolution is actually a thorny problem. First, because of weak hardware and poor environment, the images captured from digital devices are originally corrupted with motion blurs and optical distortions. Thus, SR methods have to take into account these factors and give the reasonable solu tions. Second, reconstructing the high resolution image from

a low resolution one is an inverse problem which means that there are multiple possibilities for final results. Therefore, SR methods should produce the most plausible super-resolved image accu rately. Third, SISR usually requires SR methods to enlarge the input LR image with tiny size to the desired size consistent with HR image. When the upscaling factor is large such as 4× and 8×, the images degraded severely and lost many useful contents, increasing the difficulty of the super-resolution process. SR methods should estimate the missing pixel values while enlarging the image size. The expected SR methods must be able to address these challenges.

1.4 Contributions

In this thesis, three novel SR methods have been proposed. The main contributions of this thesis are demonstrated as follows:

First, we propose the GAN-based SR method called, Generative Adversarial Face Hallucination through Residual In Internal Dense Network (GAFH-RIDN), to per form super-resolution on face images. We design a novel architecture called Residual in Internal Dense Block (RIDB) for generator. Furthermore, we exploit the enhanced discriminator, Relativistic Average Discriminator. By incorporating an adversarial loss and the perceptual loss, our method can be optimized greatly, producing realistic high quality face images.

Second, we propose the second SR method named, Semantic Encoder guided Generative Adversarial Face Ultra-Resolution Network (SEGA-FURN). The SEGA FURN has the ability to handle large upscaling factors such as 4× and 8×. We design the semantic encoder to obtain the embedded semantics which can reflect face attributes. Moreover, we propose the joint discriminator which not only distinguishes

the image data but also determines whether the input embedded semantics is from the real HR image or fake generated image. With the help of these properties, SEGA FURN is able to alleviate the gradient vanishing problem, resulting in a stable training process. Extensive experiments on large scale datasets CelebA demonstrated that our method can produce photo-realistic super-resolution images and also greatly improve quantitative values compared with state-of-the-art methods.

Third, we concentrate on the real-world natural image super-resolution task. It is an unsupervised task where there is no paired LR-HR image data and LR images come from the real-world with complicated degradation kernels. To fulfill the real world image super-resolution task, we propose the Unsupervised Bi-directional Cycle Domain Transfer Learning based Generative Adversarial Network (UBCDT-GAN). Our method comprises Unsupervised Bi-directional Cycle Domain Transfer Network (UBCDTN) and Semantic Encoder guided Super Resolution Network (SESRN). In order to simulate the real-world LR image characteristics, we utilize UBCDTN to generate a real-like LR image which contains the similar characteristics with the real-world LR image. Then, we employ the SESRN on the generated real-like LR image and upscale it to the HR image size while recovering finer details for the super resolved image. We evaluate our method on real-world dataset, New Trends in Image Restoration and Enhancement (NTIRE) 2020 Track 1. The super-resolution results demonstrate that our method can achieve promising performance than other state of-the-art methods quantitatively and qualitatively.

1.5 Thesis Structure

This thesis comprises six chapters, including the current introduction chapter. The structure of the remaining parts of this thesis are summarized below: Chapter 1: We introduce the basic concept of single image super-resolution, mo tivation of our work as well as the current challenges and contributions which are proposed in this thesis.

Chapter 2: We present the related work of state-of-the-art super-resolution meth ods. We give the overview of current super-resolution method categories, consisting of traditional methods, and deep learning-based methods. Then, we provide the deep learning background related to the proposed method and further study the quantita tive evaluation metrics of super-resolution.

Chapter 3: We demonstrate the first proposed method, GAFH-RIDN and its com-

ponents in chapter 3 specifically. Also, we provide the training strategy and experi mental results.

Chapter 4: We present the second proposed method, SEGA-FURN. This chapter covers the whole content of SEGA-FURN, including methodology, optimization

func tion and experimental analysis.

Chapter 5: We proposed the third novel method, UBCDT-GAN. In this chapter, we aim to handle the real world image super-resolution problem. The details of UBCDT GAN including architecture, objective function and ablation study are demonstrated. Furthermore, we compare the quantitative and qualitative results with the latest methods, indicating the advantage of our method.

Chapter 6: We provide the summary of this thesis and emphasize the contributions of our works.

All the references are presented at the end of this thesis.

Chapter 2

Related Work

The single image super resolution has a long history. In this chapter, before introduc ing the details of our work, we provide the related work of super-resolution methods ranging from traditional methods to deep learning-based methods. Furthermore, since our methods involve several deep learning techniques, we give the basic background of Deep Convolutional Neural Networks (DCNNs) and Generative Adversarial Networks (GANs) which are related to our work.

2.1 Traditional Super Resolution Methods

The traditional super-resolution methods can be classified into interpolation-based methods [4, 64], reconstruction-based [61, 93] and learning-based methods [22, 69].

Interpolation-based SISR methods, such as bicubic, nearest neighbor, and bilinear, rely on mathematical techniques. The basic concept of interpolation-based method is to enlarge image size and estimate an unknown pixel by its surrounding neighbors. However, they are too simple to solve complex SISR problems when the input LR images degrade severely. The reconstruction-based methods apply the prior knowl edge and employ regularization constraint on the LR image, which can produce the final HR image. The limitation is that the estimated prior knowledge lacks strong generalization ability, resulting in poor performance for arbitrary images. The third one is learning-based method, which learns the latent mapping between LR data and HR data. By the established mapping relationships, the learning-based method can estimate new HR images from given LR images. However, the learning-based method performs unsuccessfully when the upscaling factor is large.

2.2 Deep Learning-based Super Resolution Meth ods

In recent years, deep learning techniques have developed actively and widely ap plied to the super-resolution field. There are a large variety of deep learning-based methods that have been proposed, ranging from the fundamental Deep Convolutional Neural Networks (DCNNs)-based methods to the promising Generative Adversarial Networks (GANs)-based methods. Compared to traditional methods, deep learning based methods have shown the powerful

learning ability. In the following sections, we will first introduce the deep learning backgrounds and then review the deep learning based SR methods.

2.2.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a special type of Artificial Neural Net work (ANN) containing several layers in a sequence. It has become more popular in computer vision tasks such as image classification [34, 68], image segmentation [71, 103], and image enhancement [46, 77]. CNNs are the typical feedforward hierar chical network, where several different type layers are constructed to form the basic CNNs structure, including convolutional layer, activation layer, pooling layer (also called subsampling layer) and fully connected layer. The convolution kernel in the convolutional layer has the attribute to extract useful features from local image data.

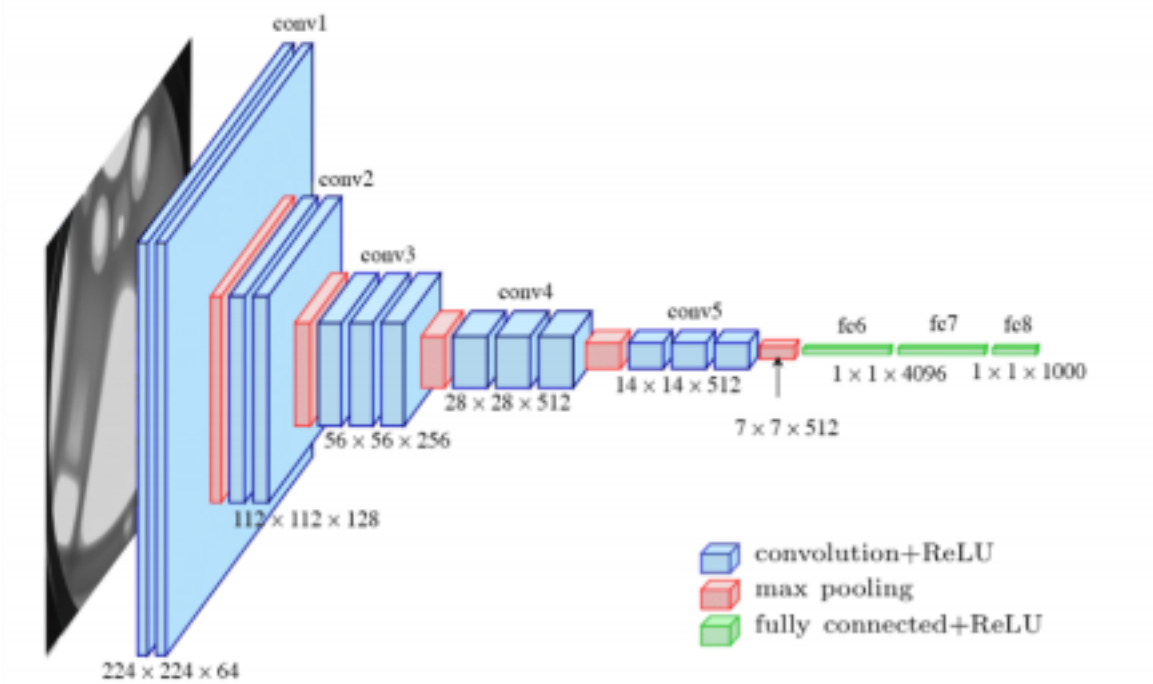

Figure 2.1: Example architecture of the Deep Convolutional Neural Networks (DC NNs) [68] for an image classification task. It includes an input layer, several convo lutional layers with activation layers, pooling layers, fully connected layers and final classification layer.

Then, the output features of the convolutional layer are passed to the non-linear

acti vation layer, which can enforce the features to be non-linear. Next, the outputs of the activation layer are usually assigned to the pooling layer, where the high dimensional features are reduced to low dimensional in order to avoid overfitting problem. Finally, the fully connected layer takes the low dimensional vector as the input and output the vector indicating the classification prediction. Figure 2.1 shows a basic CNNs architecture for an image classification task. We will present these typical layers in the following.

1) Convolutional Layer: There are a group of convolution kernels consisting of many learnable parameters. In the forward process, the convolutional layers receive the input images or the feature maps from the previous layer and then perform the convolution operation, extracting useful image features gradually. In the backward process, by employing a back propagation optimization algorithm, these kernels are able to learn desired kernel parameters. In other words, the optimal kernels are obtained, which means that they can be activated when the specific type of features

are passed into convolutional layers, so as to provide accurate feature maps for the network. The extracted feature maps can be formulated as:

$$L^i = H^i \otimes L^{i-1} + b^i \quad (2.1)$$

where L^i denotes the output of i-th convolutional layer, $L^{i-1}$is the output of the previous convolutional layer. $H^i$is i-th kernels and $b^i$is bias. $\otimes$ represents the convolution operation. The networks containing several such cascaded convolutional layers are named Deep Convolutional Neural Networks (DCNNs).

2) Activation Layer: The activation layer, also known as non-linearity layer, is an essential element in the CNNs. The outputs of the convolutional layer are passed into the activation layer. According to the correlation between the input neuron value and the network prediction value, the activation function determines whether the input value can be activated or not. In other words, the activation function acts as the ‘gate’ to select which neurons should turn on or turn off. In addition, activation functions have the attribute to normalize the input value to a range between 1 and 0 or between -1 and 1, helping CNNs to learn more complex input data.

3) Pooling Layer: The pooling layer is another important component of CNNs.

The goal of the pooling layer is to gradually reduce spatial dimensionality of the feature maps obtained from the intermediate hidden layers and relieve computation budget in the training process. Normally, there are two types of pooling opera tions: Max-pooling and Average-pooling. Max-pooling operation is integrated in max-pooling layer where it aims to choose the maximum activation pixel values from kernel covered areas of the feature maps. The average operator is another type op eration in the pooling layer. It takes the feature maps obtained by the previous convolutional layer as the input and applies the average operation to the pixels which are covered by kernels, producing the average feature maps.

4) Fully Connected Layer: The fully connected layer is an essential component of CNNs. First, the input data of CNNs is passed to convolution blocks. After a set of convolution blocks, the desired feature maps are extracted. Next, these feature maps are fed to the final fully connected layer where the 2D feature maps will be transformed to 1D vector. In the end, the 1D vector produced by a fully connected layer determines the label or category of the input data. The fully connected layer plays the critical role in image classification and recognition tasks.

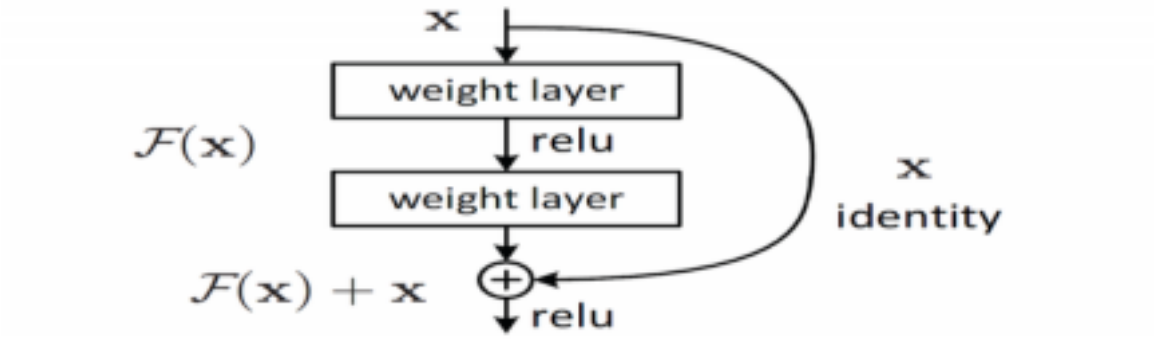

Figure 2.2: The architecture of residual block in ResNet [31].

2.2.2 ResNet and DenseNet Architectures

ResNet: The residual neural network was proposed by He [31], which achieves re markable progress in recent deep learning applications. As the DCNNs going deeper and wider, the gradient vanishing problem occurred normally, resulting in an unstable training process. He [31] proposed a deeper network called ResNet which contains 152 layers. As the dominant novelty, the proposed residual

learning further alleviates the problem of gradient vanishing and enhances training stability. In the ResNet, the residual learning is formulated in the corresponding layers, where the shortcut connections encourage that the input layer can connect with the output layer directly.

The architecture of the residual block in the ResNet is shown in Figure 2.2. The formulation of residual learning is expressed as:

$$F(x) = H(x) - x \quad (2.2)$$

where x denotes the input data, $H(x)$ is the desired underlying mapping, $F(x)$ denotes the residual function. By introducing the shortcut connections between the input and output of a layer, the identity mapping is formed. With the help of residual blocks with identity mapping, the network is able to be deeper and wider without concerning the vanishing gradient problem.

DenseNet: Densely connected convolutional networks (DenseNets) was proposed by Huang *et al*. [34], which further applies the improved shortcut connections. Unlike the skip connections in the residual block, the proposed dense connections link all succeeding layers. In other words, the input of each layer includes the features of all previous layers, and the output of each layer is propagated to all later layers. In

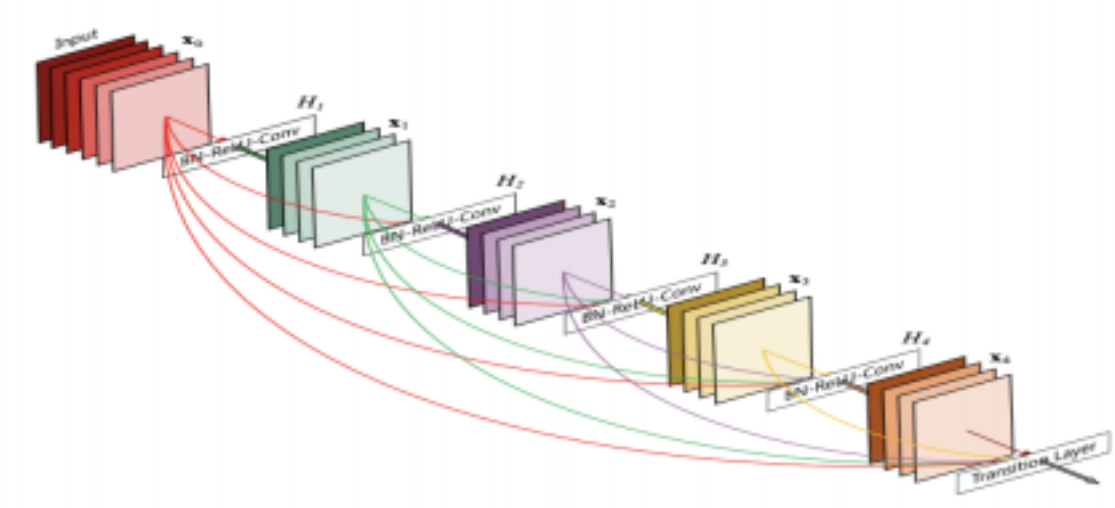

Figure 2.3: The architecture of dense blocks in DenseNet [34].

addition, DenseNet uses concatenation operations to combine features from different layers. The basic architecture of DenseNet is shown in Figure 2.3.

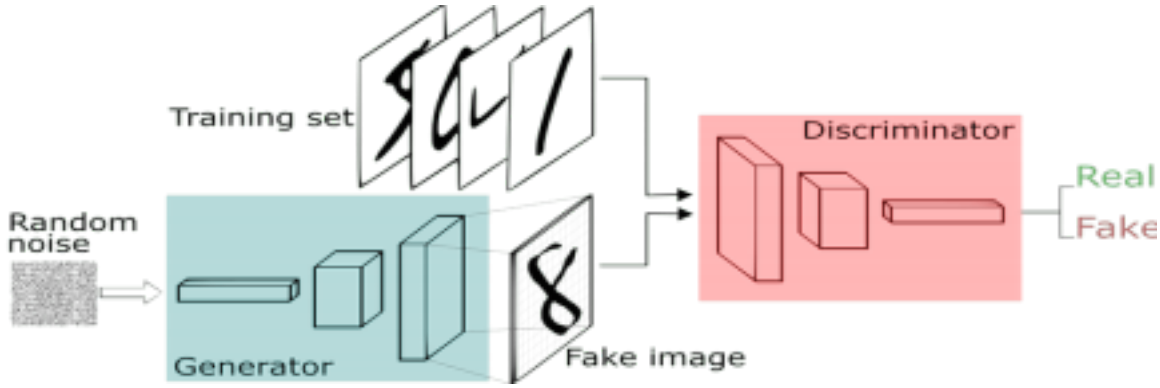

Figure 2.4: The basic pipeline of architecture of Generative Adversarial Networks (GANs) [21]

2.2.3 Generative Adversarial Networks

Generative adversarial networks (GAN) [26] is a promising advanced generative model which is widely used in unsupervised computer vision fields such as image-to-image translation [37, 74], image generation [12, 30, 49], and 3D image synthesis [24, 79]. A basic GAN consists of two main networks: the generator and the discriminator. The learning procedure of these two networks can be considered as the adversarial game

where generator and discriminator compete against each other. As shown in Figure 2.4, we give an example of GANs mechanism, where both generator and discriminator are CNN based networks. Specifically, the generator takes the random noises Z sampled from a Gaussian distribution, and produces the samples (fake image) which should be consistent with the distribution of real samples. In the adversarial training process, the discriminator receives both input data from the generator or real dataset and predicts the probability that the input samples belong to real data, while the generator attempts to produce realistic samples which can fool the discriminator. The optimization function for GANs can be expressed as follows:

$$L_{GAN} = \min_{G}\max_{D} V(D, G) = \mathrm{E}_{x\sim p_{data}(x)}[log(D(x))] + \mathrm{E}_{z\sim p_z(z)}[log(1 - D(G(z)))] \quad (2.3)$$

where $x \sim p_{data}(x)$ represents the real training data distribution and $z \sim p_z(z)$ is the Gaussian distribution. $D(\cdot)$ denotes the prediction of discriminator and $G(\cdot)$ is the generated image by generator. As Equation 2.3 shows, adversarial learning can be regarded as the min-max problem, since the generator tries to minimize the probability predicted by the discriminator while the discriminator aims to maximize the probability. When the discriminator fails to distinguish whether the input data comes from the real training dataset or the generator, $p_z(z) = p_{data}(x)$, it means that the optimal state is achieved, resulting in the desired GANs.

However, the GANs have some disadvantages. The discriminator fails to provide sufficient information for the generator to optimize, resulting in a vanishing gradient problem. Besides, the mode collapse is the most common problem during GANs training procedure, leading to unacceptable generated results. Thus, there are several GAN variants that attempt to solve these problems, such as WGAN [6], RaGAN [39], LSGAN [55]. In our proposed methods, we exploit some improved GANs to further improve model robustness.

2.2.4 Deep Convolutional Neural Networks (DCNNs)-based SR Methods

Recently, many deep learning techniques, such as DCNNs, GANs, have been applied to SISR problems. This section provides some significant methods from various as pects.

1) Pre-upsampling SISR Networks: In the earlier DCNNs-based SR methods,

since super-resolving the low resolution images to high resolution images directly is super difficult, most models utilize traditional SR interpolation methods, such as bicubic, bilinear, to upscale the LR input to the same dimension as HR image first and then learning the mapping from LR images to the desired HR images. The pioneer pre-upsampling method is SRCNN [17] proposed by Dong *et al*. Inspired from the sparse coding-based SR methods, SRCNN utilizes convolutional layers to establish the relationship between LR images and HR images. In the SRCNN, the LR images are upsampled by pre-processing step before feeding into convolutional layers. Then the enlarged LR input is passed into the SRCNN model where it consists of three convolutional layers to produce the final HR image. Kim *et al*. [42] proposed a very deep convolutional neural network, namely VDSR. Learned from the structure of VGG16 [68], the VDSR cascades 20

convolutional layers with small filters. The upsampled LR image produced by bicubic is fed into VDSR, which is the same as SRCNN.

2) Post-upsampling SISR Networks: The computational complexity and training time of pre-upsampling methods is higher. To address this problem, post upsampling methods is widely applied to SISR. The post-upsampling methods di rectly take tiny LR images as the input and then upsample the feature maps ex tracted by convolutional blocks to the desired dimensionality. The representative post-upsampling SISR network is Fast Super-Resolution Convolutional Neural Net work (FSRCNN) [18]. Instead of interpreting the input LR image in advance, FSR CNN employs the deconvolution layer to upsample the input image. Owing to this property, the FSRCNN is able to improve SISR performance and reduce the com putational cost greatly. In addition, She *et al*. [65] proposed an Efficient Sub-Pixel Convolutional Neural Network (ESPCN). In ESPCN, several convolutional blocks extract the feature representations from the input LR image. Then, it applies the designed sub-pixel convolutional layer to map the high dimensional feature vectors to the reconstructed HR images.

3) Residual Learning for SISR: Most pre-upsampling and post-upsampling methods are shallow networks, which have a straightforward architecture with small kernel size, ignoring a large amount of valuable feature representations from the input LR image. One milestone contributions in SISR is that residual learning is involved to build the deeper SISR models, which aims to learn the residual mapping between input LR images and HR images. In addition, residual learning with skip connections can greatly alleviate vanishing gradients and make use of fully feature representations.

Inspired from the ResNet [31] architecture, Lim *et al*. [48] proposed the Enhanced Deep Super-Resolution (EDSR), which is the deeper DCNNs model containing 32 con volutional layers. A number of residual blocks employed in EDSR can make EDSR easily goes to deeper and stable training process. Another typical method is Cascad ing Residual Network (CARN) [3] proposed by Ahn *et al*., which designs local and global cascading modules based on a residual network. By further exploring residual learning, CARN is able to utilize multi-level features to recover image details so as to improve resolution of the input LR image. Overall, the residual learning is capable of producing desired super-resolution results

4) Densely Connected Networks for SISR: Motivated by the development of the DenseNet [34], a large number of SR methods based on densely connected mechanism have been proposed to enhance SISR performances. Most recent work is SRDenseNet [72] where the dense connections are employed to link all the layers. Since the dense skip connections can connect the input layer and the final upsampling layer, it is able to make use of different level features, providing enough feature information for the reconstruction process. Furthermore, by combining the residual skip connections and dense short connections, Zhang *et al*. [98] proposed Residual Dense Network for SISR (RDN). In RDN, the new Residual Dense Block (RDB) is introduced, where it includes local residual learning and global residual learning. At the same idea, Haris *et al*. [29] proposed Deep Back-Projection Networks (DBPN), which utilizes the dense connection to link the upsampling and downsampling layers. Overall, a group of SR methods with densely connected architecture are able to take advantage of multiple level features to improve SR performances greatly.

2.2.5 Generative Adversarial Networks (GANs)-based SR Meth ods

Before introducing GAN-based SR methods, we first present the drawback of DCNNs based methods. As described in the above sections, all of these methods utilize Mean Squared Error (MSE) as the objective function to minimize the loss between SR im ages and HR images. The MSE loss is a straight forward optimization strategy and simple to implement. It is widely applied to image processing, pattern recognition tasks. However, MSE has some severe problems, where the results optimized by MSE are unfaithful to human perception, failing to provide visually pleasant images. The reason is that MSE loss assumes that there is no correlation between the influence of

noise and the characteristics of an image, which is inconsistent with many application settings. By contrast, the human perception is more sensitive to these delicate char acteristics and structures. Thus, in order to eliminate the drawbacks of MSE loss and comply with human visual perception, the perceptual loss and adversarial loss are introduced to handle this problem. The perceptual loss is the feature-based objective loss function which is calculated on feature maps extracted from pre-trained DCNNs. The term of adversarial loss is introduced by GANs, where the adversarial loss is able to guide the network to be optimal and

produce realistic super-resolution images.

Recently, GANs have become more popular in the computer vision field, and widely used in the SISR area. Ledig [46] proposed SRGAN, which is the first gener ative adversarial network for SISR task, including the generator and discriminator. The SRGAN employs the novelty learning strategy comprising the perceptual loss and an adversarial loss. With the help of these two losses, the SRGAN can improve the visual effect of generated SR images. Furthermore, Wang *et al*. [77] proposed Enhanced SRGAN (ESRGAN). They strengthen the SRGAN from three aspects: network architecture, discriminator and objective loss. Eventually, the ESRGAN generates better desired images compared to SRGAN. In [59], Park proposed SRFeat based on the GAN framework. According to perceptual loss and pixel-wise loss, SR Feat can optimize the training process and produce fine results with high-frequency details. Owing to an effective adversarial learning strategy, most recent SR methods resort to GANs for handling SR tasks.

2.2.6 Quantitative Evaluation Metrics for Super Resolution Methods

Image quantitative evaluation metrics are significant for super-resolution tasks. The appropriate metrics provide an effective way to measure the quality of SR images. The common and classic methods are: Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [78]. In our experiments, we utilize these two metrics to evaluate the SR images generated by our proposed methods. Also, we use both two to compare our methods with state-of-the-art methods. Both two classic image evaluation methods are introduced in the following.

1)PSNR: It is the most common and widespread metric in the image processing field. Before calculating PSNR, we first define the Mean Squared Error (MSE). The

formulation for MSE is:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I_{HR}(i, j) - I_{SR}(i, j)]^2 \quad (2.4)$$

where I_{HR} is the real HR image with size $m \times n$ and I_{SR} is the SR image generated by the method. $I_{HR}(i, j)$ and $I_{SR}(i, j)$ represent the pixel positions in I_{HR} and I_{SR} respectively. According to the MSE, the similarity of I_{HR} and I_{SR} can be reflected. Based on MSE, the PSNR can be formulated as:

$$PSNR = 10 \times log_{10}[(MAX_I - MIN_I)^2 \quad MSE] \quad (2.5)$$

where MAX_I and MIN_I represent the highest and lowest pixel values of image I. Commonly, this formulation merely measures one channel images, such as gray im ages. To evaluate the RGB images containing three channels, we have to estimate PSNR for each channel (Red, Green, Blue) first, and then calculate the average of them. The higher PSNR value indicates the better image quality.

2) SSIM: Structural Similarity is another well-known image quality method. It is a perception-based method, where SSIM takes into account three important independent components of the input image: Luminance, Contrast and Structure. Specifically, these factors can reflect the structural information and the degree of pixel inter-dependencies. Here, we define $x_i$is I_{HR} and y_i equals to I_{SR}. The SSIM of two images x_i and y_i can be denoted as:

$$SSIM(x, y) = [l(x, y)]^{\alpha} \cdot [c(x, y)]^{\beta} \cdot [s(x, y)]^{\gamma} \quad (2.6)$$

where l, c, s denotes the Luminance, Contrast, Structure of x_i and y_i respectively. The Luminance $l(x, y)$ is calculated through the mean intensity, which is expressed as:

$$\mu_x = \frac{1}{N} \sum_{i=1}^{n} x_i, \quad \mu_y = \frac{1}{N} \sum_{i=1}^{n} y_i \quad (2.7)$$

and the Contrast $c(x, y)$ is evaluated by eliminating the mean intensity from the image and then taking the standard deviation, which is represented as:

$$c(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu^2_x + \mu^2_y + C_1} \quad (2.8)$$

where C_1 and C_2 are two constants. And they can be defined as: $C_1 = (K_1L)^2$, $C_2 =$

$$(K_2L)^2 \quad (2.9)$$

where L is the range of the pixel values of the input image, and normally K_1 and K_2 are equal to 0.01 and 0.03. As for the Structure $s(x, y)$, it can be calculated through taking normalization of standard deviation. The formulation is expressed as:

$$\sigma_x = \left[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)\right], \quad \sigma_y = \left[\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu_y)\right] \quad (2.10)$$

Finally, we combine all the three factors to construct the SSIM formulation. There are three parameters: α, β and γ in Equation 2.6. We simply set these three parameters to 1 and the SSIM can be expressed as:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu^2_x + \mu^2_y + C_1)(\sigma^2_x + \sigma^2_y + C_2)} \quad (2.11)$$

Overall, the SSIM metric can reflect the image quality perceptually.

Chapter 3

End-to-End Generative Adversarial Face Hallucination through Residual In Internal Dense Network

3.1 Overview

Face hallucination has been a highly attractive computer vision research topic in re cent years. It is still a particularly challenging task since the human face has a complex and delicate structure. In this chapter, we propose a novel network structure, namely end-to-end Generative Adversarial Face Hallucination through Residual in Internal Dense Network (GAFH-RIDN), to hallucinate an unaligned tiny (32×32 pixels) low resolution face image to its 8× (256×256 pixels) high-resolution counterpart. We propose a new architecture called Residual in Internal Dense Block (RIDB) for the generator and exploit an improved discriminator, Relativistic average Discriminator (RaD). In GAFH-RIDN, the generator is used to generate visually pleasant hallu cinated face images, while the improved discriminator aims to evaluate how much input images are realistic. With continual adversarial learning, GAFH-RIDN is able to hallucinate perceptually plausible face images. Extensive experiments on large face datasets demonstrate that the proposed method significantly outperforms other state-of-the-art methods.

3.2 Introduction

Face Hallucination (FH), also known as Face Super-Resolution (FSR), is a domain specific image Super-Resolution (SR) problem, which refers to hallucinate the High Resolution (HR) face images from their Low-Resolution (LR) counterparts. It is a significant task in the face analysis field, which is of remarkable benefit to computer vision applications such as face surveillance [105] and recognition [83]. However, face hallucination is an ill-posed inverse problem and particularly challenging since the LR image may correspond to many HR candidate images and has lost many crucial facial structures and components [13, 87, 89]. In order to hallucinate high quality face images, many FH methods have been proposed. Generally, we can classify these approaches into two categories: traditional methods and deep learning-based methods.

Many traditional methods have been proposed to address face hallucination tasks [7, 76, 82]. Baker and Kanade [7] presented the image pyramid model to learn the best relationship between LR and HR patches, which can reconstruct high-frequency details of LR face images. In [76], Wang and Tang employed eigen-transformation to build a linear mapping between LR and HR face subspaces. By adopting relationship

between particular facial components, Yang *et al*. [82] combined the face priors to recover facial information from HR image components.

Recently, deep learning-based methods have emerged and achieved the state-of the-art performance [17, 42, 100, 102]. Dong *et al.* [17] firstly introduced a deep learning-based SR method named SRCNN that directly learned an end-to-end map ping between HR images and LR images. In [100], Zhou *et al.* presented the novel Bi channel convolutional network to hallucinate face images in the wild. The Cascaded Bi-Networks (CBN) was presented by Zhu *et al.* [102], in which two sub-networks (face hallucination and dense correspondence field estimation) were optimized alternately.

The limitation of the above face hallucination methods is that they utilize re construction loss such as $L1$ or $L2$ to optimize the hallucination process, which is prone to producing over-smoothed hallucinated images even though these models obtained higher Peak Signal-to-Noise Ratio (PSNR) value [78]. To

address this prob lem, several Generative Adversarial Network (GAN) -based models were proposed [13, 46, 77, 87, 88, 90]. It is proved that GAN-based models using powerful constraint losses are able to further generate visually realistic HR images [95]. Christian *et al.*'s work [46] extended GAN to the SR field and proposed an effective method, called SR GAN utilizing an adversarial loss and the perceptual loss. Following SRGAN, Wang *et al.* [77] presented the ESRGAN by proposing a new generator architecture and using improved perceptual loss. Yu and Porikli [91] proposed MTDN based on GAN. Nevertheless, when the input resolution is super low, it fails to recover high quality face images, leading to blurred patterns and severe artifacts.

However, the aforementioned GAN-based face hallucination models are prone to model collapse [13, 46, 77, 88], resulting in ghosting artifacts in the hallucinated results, especially when the input image resolution is extremely low. To address this problem, in this chapter, we propose a novel GAN-based FH method, end-to end Generative Adversarial Face Hallucination through Residual in Internal Dense Network (GAFH-RIDN), as shown in Figure 3.1. The contributions of this chapter are mainly in four aspects:

1) Our proposed method is capable of hallucinating an LR (32×32 pixels) un aligned tiny face image to a Hallucinated Face (HF) image (256×256 pixels) with an ultra upscaling factor 8×.

2) We propose the Residual in Internal Dense Block (RIDB), which boosts the flow of features through the generator and provides hierarchical features for the hal lucination process.

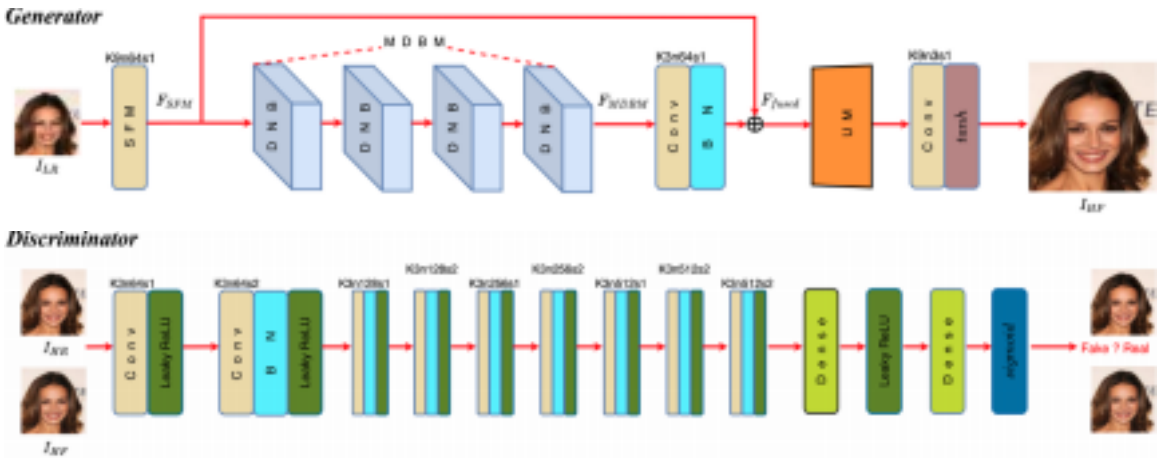

Figure 3.1: The architecture of our end-to-end Generative Adversarial Face Halluci nation through Residual in Internal Dense Network (GAFH-RIDN). I_{HF}

represents HF image. I_{HR} and I_{LR} denote HR and LR face image respectively. K, n, and s represent kernel size, the number of feature maps and strides respectively. SFM is the Shallow Feature Module. MDBM describes the Multi-level Dense Block Module. UM is the Upsampling Module. DNB represents the Dense Nested Block as shown in Figure 3.2.

3) We exploit the Relativistic average Discriminator (RaD) [39], which evaluates the probability that the given HR face images are more realistic than HF images.
4) Contrary to classical face hallucination methods [13, 53, 87], our method does not involve any prior information or claim facial landmark points for its hallucinating, which facilitates the whole training process and enhances the model robustness.

3.3 Proposed Method

In this section, we will first describe the proposed architecture and demonstrate the Residual in Internal Dense Block (RIDB). Next, we will discuss the improved discrim inator. Finally, we will present the perceptual and adversarial losses function used in the GAFH-RIDN. The architecture of GAFH-RIDN is shown in Figure 3.1.

3.3.1 Network Architecture

As shown at the top of Figure 3.1, the proposed generator mainly consists of three stages: Shallow Feature Module (SFM), Multi-level Dense Block Module (MDBM),

and Upsampling Module (UM). The LR face image I_{LR} is fed into the SFM as the initial input. At the end, hallucinated face image I_{HF} is obtained from the UM. As for the SFM, we utilize one convolutional (Conv) layer to extract the shallow feature maps. It can be expressed as follows:

$$F_{SFM} = f_{Conv}(I_{LR}) \quad (3.1)$$

where f_{Conv} represents the Conv operation in the SFM. F_{SFM} denotes the shallow (low-level) features and serves as the input to the MDBM. The following module MDBM is built up by multiple Dense Nested Blocks (DNBs) formed by several

RIDBs, which will be discussed in the next subsection. The procedure of high-level feature extraction in MDBM can be formulated as:

$$F_{MDBM} = f_{DNB,i}(f_{DNB,i-1}(\cdots(f_{DNB,1}(F_{SFM}))\cdots)) \quad (3.2)$$

where $f_{DNB,i}$ denotes high-level feature extraction of the i-th DNB, F_{MDBM} represents the high-level feature extracted by MDBM. As for each DNB, it includes 3 RIDBs cascaded by residual connections and one scale layer, as shown in Figure 3.2. It can be formulated as:

$$F_{DNB,i} = \alpha F_{i,j}(F_{i,j-1}(\cdots F_{i,1}(F_{DNB,i-1})\cdots)) + F_{DNB,i-1} \quad (3.3)$$

where $F_{DNB,i-1}$, $F_{DNB,i}$ denotes the input and output of i-th DNB, $F_{i,j}$ represents the j-th RIDB of the i-th DNB. We empirically assign α to be 0.2 in the scale layer. Next, the low-level and high-level features should be fused to boost hallucination performance via skip connection. Let F_{fused} denotes the fused feature, the feature fusion process can be expressed as:

$$F_{fused} = f_{Conv}(F_{MDBM}) + F_{SFM} \quad (3.4)$$

Furthermore, the fused feature F_{fused} is passed to the UM followed by one Conv layer. And then, the fused feature is transformed from the LR space to the HR space through upsampling layers in the UM. The hallucination process can be formulated as:

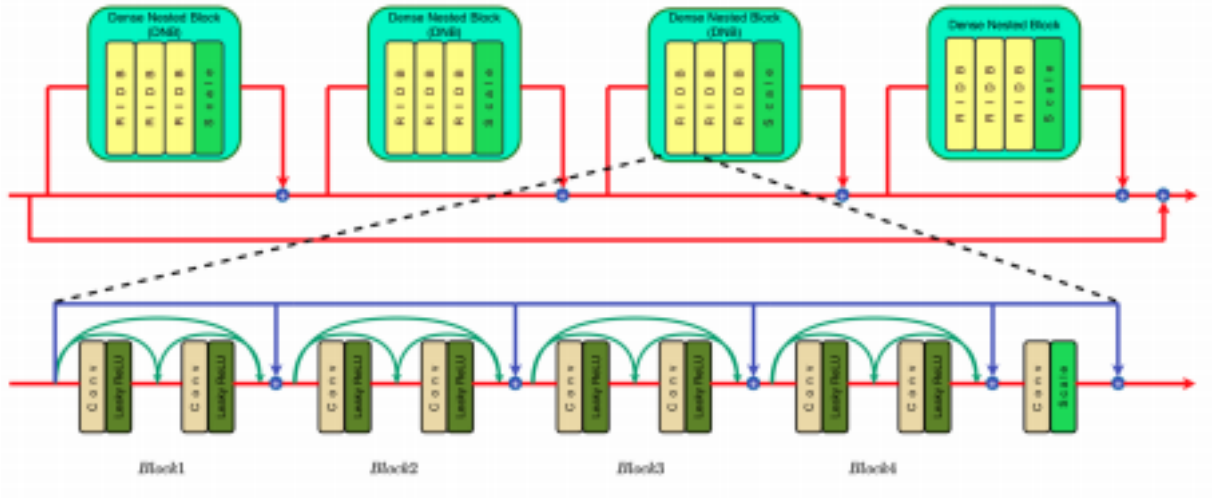

Figure 3.2: Top: Dense Nested Block (DNB) composed of multiple RIDBs. Bottom: The architecture of our proposed Residual in Internal Dense Block (RIDB).

$$I_{HF} = f_{UM}(F_{fused}) = H_{GAF\,H-RIDN}(I_{LR}) \quad (3.5)$$

where f_{UM} represents the upsampling operation in the UM, $H_{GAF\,H-RIDN}$ denotes the function of our GAFH-RIDN. Finally, we obtain the HF image I_{HF}.

3.3.2 Residual in Internal Dense Block

As mentioned in Section 3.3, we propose a novel architecture RIDB for the generator, which is used to form the DNB (as shown in Figure 3.2). The proposed RIDB is able to extract hierarchical features and address the vanishing-gradient problem, which is the commonly encountered issue in [46, 72, 77, 88, 98]. The proposed RIDB is made up of four internal dense blocks and all the internal dense blocks are cascaded through residual connections performing identity mapping. The architecture of the RIDB is expressed as:

$$F_{RIDB,p} = F_{p,q}(F_{p,q-1}(\cdots F_{p,1}(F_{RIDB,p-1})\cdots)) + F_{RIDB,p-1} \quad (3.6)$$

where $F_{RIDB,p-1}$ and $F_{RIDB,p}$ denote the input and output of the p-th RIDB re spectively, $F_{p,q}$ represents the q-th internal dense block of p-th RIDB. In addition, an internal dense block is a composition of two groups of the Conv layer followed by the LeakyReLU activation layer. And the two groups are linked by dense skip connections. Each internal dense block can be calculated as follows:

$$F_{q,k} = \delta(W_{q,k}[F_{q,k=1}, F_{q,k=2}]) \quad (3.7)$$

where $F_{q,k}$ represents the output of the k-th Conv layer of q-th internal dense block. $[F_{q,k=1}, F_{q,k=2}]$ refers to the concatenation of feature maps in q-th internal dense block. $W_{q,k}$ is the weights of the k-th Conv layer. δ denotes the LeakyReLU activation. By involving residual learning and more dense connections in the RIDB, the feature maps of each layer are propagated into all succeeding layers, promoting an effective way to extract hierarchical features. Thus, our proposed method is capable of obtain ing abundant hierarchical feature information and

alleviating the vanishing-gradient problem.

3.3.3 Improved Discriminator

Instead of using the discriminator of Standard GAN (SGAN) [26], inspired by [39], we adopt the Relativistic average Discriminator (RaD) in our method. Thanks to RaD, the discriminator of GAFH-RIDN has the ability to distinguish how the given HR face image is more authentic than the hallucinated face image. The architecture of our discriminator is shown at the bottom of Figure 3.1. The limitation of the SGAN in [26, 46, 88] is that they only concentrate on increasing the probability that fake samples belong to real rather than decreasing the probability that real samples belong to real simultaneously. In other words, the standard discriminator ignores real samples during the learning procedure [39]. As a result, the model can not provide sufficient gradients when updating the generator, which causes the problem of gradient vanishing for training the generator. The standard discriminator can be expressed as:

$$D(x) = \sigma(C(x)) \quad (3.8)$$

where x can be either I_{HR} or I_{HF} in this context, σ represents the sigmoid function, and $C(x)$ denotes the output of a non-transformed discriminator. As Equation 3.8 shows, the standard discriminator only evaluates the probability for a given real sample or a generated sample. According to [39], RaD takes into con sideration how a given real sample is more authentic than a given generated sample. The RaD can be formulated as:

$$D(x_r, x_f) = \sigma(C(x_r) - E_{xf}[C(x_f)]) \quad (3.9)$$

where E_{xf} denotes the average of the fake samples in one batch. Contrary to standard discriminator, as Equation 3.9 shows, the probability predicted by RaD relies on both real sample x_r and fake sample x_f, which is capable of making discriminator relativistic. In our GAFH-RIDN, we can optimize the RaD by L^{adv}_{D} based on Equation 3.10, and the generator is updated by L^{adv}_{G}, as in Equation 3.11.

$$L^{adv}$$

$$L^{adv}_{D} = -E_{I_{HR}} \sim p_{(I_{HR})}[log\,(D(I_{HR}, I_{HF}))] - E_{I_{HF}} \sim p_{(I_{HF})}[log\,(1 - D\,(I_{HF}, I_{HR}))] \quad (3.10)$$

$$L^{adv}_{G} = -E_{I_{HR}} \sim p_{(I_{HR})}[log\,(1 - D(I_{HR}, I_{HF}))] - E_{I_{HF}} \sim p_{(I_{HF})}[log\,(D\,(I_{HF}, I_{HR}))] \quad (3.11)$$

where I_{HR} and I_{HF} denote HR images and HF images respectively, $D(\cdot)$ describes the probability predicted by RaD, E represents the expectation, $I_{HR} \sim P_{I_{HR}}$ and $I_{HF} \sim P_{I_{HF}}$ represents the HR images distribution and HF images distribution re spectively. Because of this property, our proposed GAFH-RIDN is capable of allowing the probability of I_{HR} being real to decrease while letting the probability of I_{HF} being real increase and benefiting from gradients of both I_{HR} and I_{HF} in the adversarial training. Therefore, our proposed method can address the gradient vanishing prob lem. Our discriminator contains 9 Conv layers with the number of 3x3 kernels and the stride of 1 or 2 alternately. The channels of feature maps increase by a factor 2, from 64 to 512. The resulting 512 feature maps are passed through two dense layers. Finally, after the sigmoid activation layer, RaD estimates the probability that the given HR face images are more realistic than HF images.

3.3.4 Perceptual Loss

We adopt the pre-trained VGG-19 [68] as the feature extractor φ to obtain feature representation used to calculate $L_{perceptual}$. We extract low-level feature maps of HR and HF images obtained by the 3^{rd} Conv layer before the 4^{th} maxpooling layer respectively. HR and HF feature maps are defined as $\varphi_{3,4}$. $L_{perceptual}$ is defined as follows:

$$L_{perceptual} = k\varphi_{3,4}(I_{HR}) - \varphi_{3,4}(I_{HF})k^2 \quad (3.12)$$

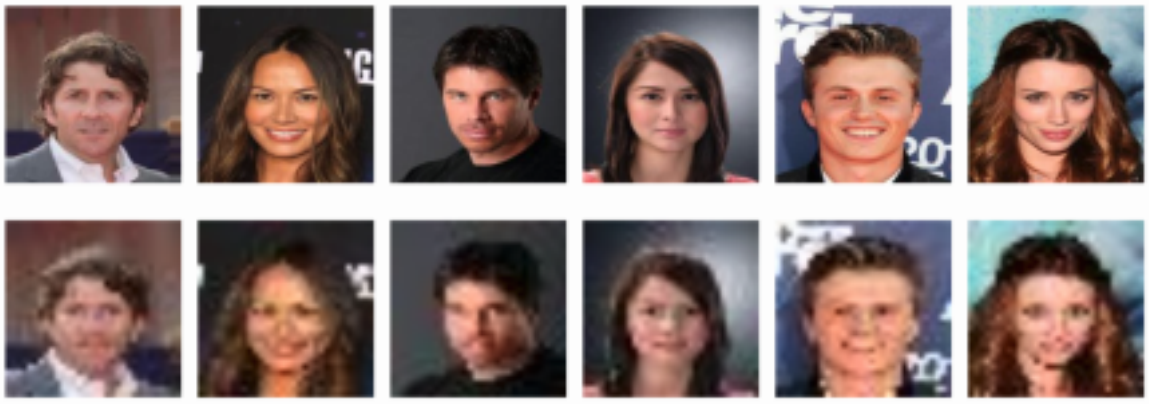

Figure 3.3: The sample images of CelebA dataset. The top row presents HR images (256×256 pixels) and the bottom row shows corresponding LR images (32×32 pixels)

3.3.5 Total Loss

The total loss function L_{total} for generator can be represented as two parts: $L_{perceptual}$ and L^{adv}_{G}. We introduce the perceptual loss to enhance perceptual quality of the HF image from the visual aspect. In addition, an adversarial loss is expected to improve the fidelity of the HF image. The formula is defined as follows:

$$L_{total} = \lambda_1 L_{perceptual} + \lambda_2 L^{adv}_{G} \quad (3.13)$$

where λ_1 and λ_2 are corresponding hyper-parameters used to balance $L_{perceptual}$ and L^{adv}_{G}. We empirically set $\lambda_1 = 1$, $\lambda_2 = 10^{-3}$ respectively.

3.4 Experiments

In this section, we will first present the details of datasets and implementation. Next, we will discuss the comparisons with the state-of-the-art methods [42, 46, 77, 86, 89, 102] qualitatively and quantitatively.

3.4.1 Implementation Details

We conducted experiments on the large-scale face image dataset, CelebFaces At tributes Dataset (CelebA) [52]. It consists of 202,599 face images of 10,177

celebrities. As shown in Figure 3.3, we give several sample images of CelebA dataset including HR face images (256×256 pixels) and its LR face versions (32×32 pixels). We randomly

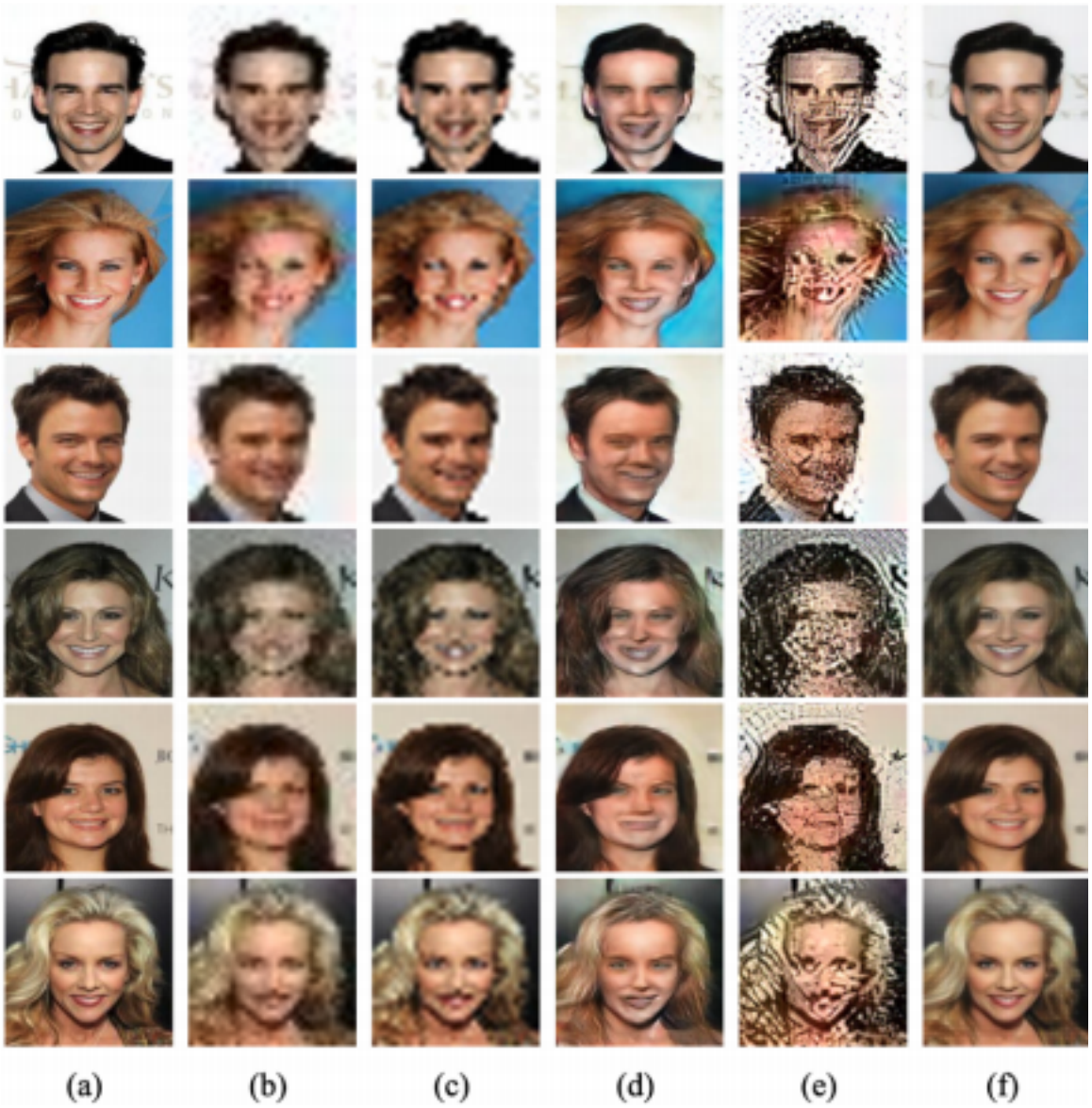

Figure 3.4: Comparison of visual results with state-of-the-art methods on scaling factor 8x. (a) HR images, (b) LR inputs, (c) Bicubic interpolation, (d) Results of SRGAN [46], (e) Results of ESRGAN [77], and (f) Our results

selected 162,048 HR face images as the training set, and the next 40,511 images were used as the testing set. We cropped the HR face images and resized them to 256×256 pixels, and then obtained LR (32×32 pixels) input images by

downsampling HR im ages using bicubic interpolation with a downsampling factor of 8×. In the proposed generator, we set the number of DNBs to 4, and totally 12 RIDBs were used. In the training phase, we trained the proposed method for 10000 epochs and the training

Method CelebA 8×	PSNR SSIM
Bicubic	22.90 0.65
VDSR [42]	19.58 0.57
CBN [102]	18.77 0.54
SRGAN [46]	20.64 0.62
FSRFCH [86]	23.14 0.68
TDN [89]	22.66 0.66
Kim [40]	22.96 0.69
ESRGAN [77]	20.32 0.57
Ours	24.28 0.71

Table 3.1: Quantitative comparison on CelebA dataset for scaling factor 8x, in terms of average PSNR(dB) and SSIM. Numbers in bold are the best evaluation results among state-of-the-art methods.

batch size was set to 8. We used Adam with $\beta_1 = 0.9$, $\beta_2 = 0.999$ to optimize the proposed method. The learning rate was set to 10^{-4}. We alternately updated the generator and discriminator.

3.4.2 Qualitative Comparison

Qualitative results among these methods are shown in Figure 3.4. We observe that the bicubic interpolation produces heavy blur and fails to generate clear textures. For SRGAN [46], it outputs noticeable artifacts around facial components, especially in the eyes, nose, and mouth regions. In particular, ESRGAN [77] produces unrealistic textures and involves severe ghosting artifacts. In contrast, it is obvious that our proposed method is capable of producing visually pleasant and authentic HF images.

3.4.3 Quantitative Comparison

Table 3.1 shows the quantitative comparison on 8× HF images. The results demon strate that our proposed method achieves the best performance among

all methods. In Particular, our method produces the highest score of 24.28dB/0.71 in terms of PSNR and SSIM respectively. Furthermore, compared with the second-best FSRFCH [86] 23.14dB/0.68, our method outperforms it with a large margin of 1.14dB/0.03. The performance proves the effectiveness of the proposed RIDB and the optimized RaD used in our method.

30

3.5 Conclusions

In this paper, we proposed a novel end-to-end face hallucination method (GAFH RIDN) to hallucinate a tiny LR (32×32 pixels) unaligned face image to its 8× HR (256×256 pixels) version. By exploiting Residual in Internal Dense Block (RIDB) and Relativistic average Discriminator (RaD), our method successfully produced photo realistic hallucinated face images. Extensive experiments demonstrated that GAFH RIDN was superior to the state-of-the-art methods on the face benchmark qualita tively and quantitatively.

31

Chapter 4

Semantic Encoder Guided Generative Adversarial Face Ultra-Resolution Network

4.1 Overview

To strengthen the robustness of previous proposed GAFH-RIDN and enhance the super-resolution performance, in this chapter, we extend the previous work (GAFH RIDN) from several aspects and propose a novel face super-resolution method, namely Semantic Encoder guided Generative Adversarial Face Ultra-Resolution Network (SEGA

FURN) to ultra-resolve an unaligned tiny LR face image to its HR counterpart with multiple ultra-upscaling factors (e.g., 4× and 8×). In our method, the proposed se mantic encoder has the ability to capture the embedded semantics to guide adversarial learning. We still utilized the proposed Residual in Internal Dense Block (RIDB) for SEGA-FURN, since it is able to extract hierarchical features for

the generator. More over, we propose a joint discriminator which not only discriminates image data but also discriminates embedded semantics, learning the joint probability distribution of the image space and latent space, and we use a Relativistic average Least Squares loss (RaLS) as the adversarial loss, which can alleviate the gradient vanishing problem and enhance the stability of the training procedure. According to extensive experi ments on large face datasets, it is obvious that our proposed method achieves superior super-resolution results and significantly outperforms other state-of-the-art methods in both qualitative and quantitative comparisons.

4.2 Introduction

Face Super-Resolution (FSR) has been a promising computer vision topic in recent years. It is widely applied to the face applications such as face surveillance [105] and identification [83]. However, there are many particular challenges in the research, one of which is that face have complex geometric structures and facial expressions and

another is that most of the face information has been lost when the resolution of the LR image is quite low. To address these problems, many FSR methods have been proposed. Generally, these methods can be divided into two categories: conventional methods and deep learning-based methods. The details of these methods can be referred to chapter 2.

As we discussed in the section 2.2.5, the latest GAN-based methods suffer from the problem of model collapse and training instability, leading to notorious oversmooth ing artifacts [46, 77]. To breakthrough the limitation of previous SR methods and produce photo-realistic SR face images, we propose a novel GAN-based SR method, namely Semantic Encoder guided Generative Adversarial Face Ultra-Resolution Net work (SEGA-FURN), as shown in Figure 4.1. The main contributions of this chapter can be summarized as follows:

1) Our proposed method is able to ultra-resolve an unaligned tiny face image to a Super-Resolved (SR) face image with multiple upscaling factors (e.g., 4× and 8×). In addition, our method does not need any prior information or facial landmark points.

2) We design a semantic encoder to reverse the image information back to

the embedded semantics reflecting facial semantic attributes. The embedded semantics combined with image data are fed into a joint discriminator. Such innovation can let the semantic encoder guide the discriminative process, which is beneficial to enhance the discriminative ability of the discriminator.

3) We utilize the previous proposed Residual in Internal Dense Block (RIDB) as the basic architecture for the generator. This innovation provides an effective way to take advantage of hierarchical features, resulting in increased feature extraction capability of the SEGA-FURN.

4) We propose a joint discriminator which is capable of learning the joint probabil ity constructed by embedded semantics and visual information (HR and LR images), resulting in a powerful discriminative ability. Furthermore, in order to remedy the problem of vanishing gradient and improve the model stability, we make use of RaLS objective loss to optimize the training process.

4.3 Related Work

In this section, we first analyze the major problems associated with the Standard Generative Adversarial Network (SGAN) [26]. Second, we present improved GAN variants which can address current problems.

4.3.1 Problem Analysis

The SGAN consists of two networks, one of which is the Generator *G*, and the other is the discriminator *D*. SGAN has been applied to many applications, such as super resolution [46], image translation [101] and face aging [5]. Through adversarial learn ing in the SGAN, the generator and discriminator compete against each other. Both two networks try to optimize themselves to solve the adversarial max-min problem. The objective function is:

$$V(G, D) = \max_{d}\min_{g} \mathrm{E}_{x\sim p_{data}(x)}[logD_d(x)] + \mathrm{E}_{z\sim pz(z)}[log(1 - D_d(G_g(z)))] \quad (4.1)$$

where $V(G, D)$ is a binary cross entropy loss which is commonly used in GAN appli cations, G is supposed to map random noise z from the prior distribution

$p_z(z)$ over real data x, and D is the discriminator which distinguishes whether its input comes from the G or real data distribution $p_{data}(x)$. The ultimate goal of SGAN is that D and G is capable of reaching Nash equilibrium state, in which once SGAN attains Nash equilibrium, the generator can generate realistic-looking images which fool the discriminator.

However, SGAN in [26, 46, 88] encounters the problem of gradient vanishing, model collapse and poor quality of the generated images. Several works [6, 39, 55] have proved that the objective function of the SGAN causes vanishing gradients, resulting in the instability of GAN training. The discriminator of the SGAN can be expressed as Equation 4.2:

$$D(x) = \sigma(C(x)) \quad (4.2)$$

where x expresses either I^{HR} or I^{SR} in this context, σ represents the sigmoid func tion, $C(x)$ is the probability predicted by the non-transformed discriminator. The restriction of the SGAN is that they only concentrate on increasing the probabil ity that fake samples belong to real rather than decreasing the probability that real samples belong to real simultaneously. In the SGAN, if the optimal discriminator is reached, it will stop learning the real data but will only focus on the fake samples. As a result, the generator cannot receive enough gradient information from real data to make progress and the authenticity of fake samples will no longer be improved. To address this problem, several improved GAN variants have been proposed to find the objective function with smoother and non-vanishing gradients.

4.3.2 Review of the Relativistic Aaverage GAN

RaGAN [39] was proposed as an improved SGAN by drawing up Relativistic average Discriminator (RaD). The RaD demonstrates that the SGAN ignores the relative discriminant information between real samples and fake samples. This key property is complemented in RaD, in which RaD not only improves the probability that the generated samples are real but also decreases the possibility that the real samples are real. The RaD can be expressed as:

$$D(x_r, x_f) = \sigma(C(x_r) - E_{xf}[C(x_f)]) \quad (4.3)$$

where E_{xf} denotes the average of the fake samples in one batch predicted by RaD. As Equation 4.3 shows, the RaD is capable of evaluating the probability

that the real image is more realistic than a fake image, which addresses the issue of vanishing gradient and improves the stability of GAN.

4.3.3 Review of the Least Squares GAN

The Least Squares GAN (LSGAN) [55] indicated that the vanishing gradient problem is mainly caused by the discriminator of SGAN using the sigmoid cross entropy loss. This work argued that the original discriminator penalizes a small decision error to update the generator which makes the generated samples stay on the correct side of the decision boundary, but are still far from corresponding real samples, leading to vanishing gradient problem during the adversarial process. Motivated by this issue, LSGAN proposed the least squares loss function to penalize large errors coming from fake samples that lie far away from the decision boundary. The formulation can be expressed as:

$$L^{LSGAN}_{D} = E_{x_r \sim p_r}[(C_d(x_{real}) - 0)^2] + E_{x_f \sim p_f}[(C_d(x_{fake}) - 1)^2] \quad (4.4)$$

Thus, the discriminator utilizing least squares loss is capable of providing sufficient gradients when optimizing the generator, which is able to remedy the vanishing gra dient problem.

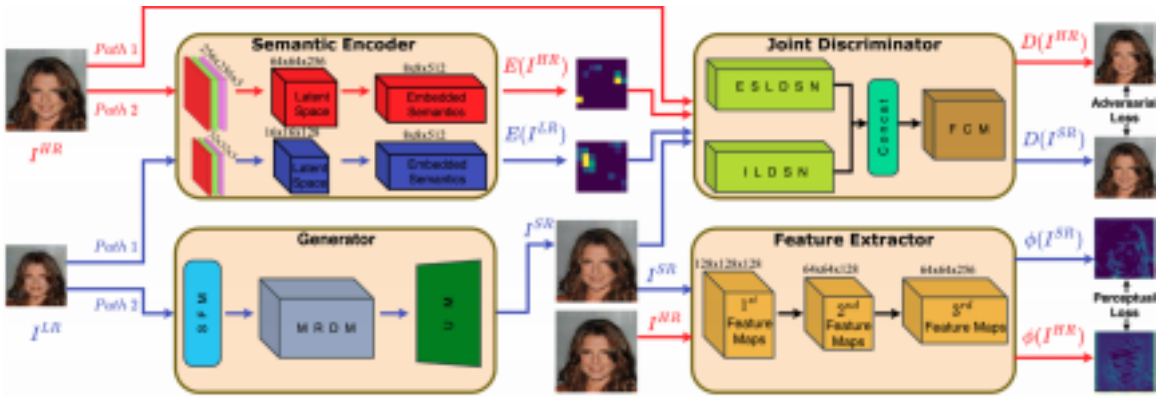

Figure 4.1: Proposed SEGA-FURN and its components: Semantic Encoder E, Generator G, Joint Discriminator D and Feature Extractor φ. For D, ESLDSN represents the Embedded Semantics-Level Discriminative Sub-Net, ILDSN represents the Image-Level Discriminative Sub-Net, and FCM denotes Fully Con nected Module. As for the generator G, there are three stages: Shallow Feature Module (SFM), Multi-level Residual Dense Module (MRDM), and Upsampling Mod ule (UM). I^{HR} and I^{LR} denote HR face images and LR face images respectively. I^{SR} is SR images from G. Furthermore, $E(\cdot)$ denotes the embedded

semantics obtained from E. $D(\cdot)$ represents the output probability of D. $\varphi(I^{HR})$ and $\varphi(I^{SR})$ describe the features learned by φ.

4.4 Proposed Method

In this section, we present the proposed method SEGA-FURN in detail. First, we describe the novel architecture of SEGA-FURN through four main parts: genera tor, semantic encoder, joint discriminator and feature extractor. Next, we introduce the objective loss function RaLS [39] for optimizing the generator and discriminator respectively. Finally, we provide the overall perceptual loss used in SEGA-FURN. The overview of our method is shown in Figure 4.1. In addition, the architecture of discriminator and generator can be seen in Figure 4.2. Moreover, the structures of the proposed DNB and RIDB are presented in Figure 4.3.

4.4.1 Generator

As shown at the top of Figure 4.2, the proposed generator mainly consists of three stages: Shallow Feature Module (SFM), Multi-level Dense Block Module (MDBM), and Upsampling Module (UM). The LR face image I^{LR} is fed into the SFM as the initial input. At the end, SR face image I^{SR} is obtained from the UM. As for the

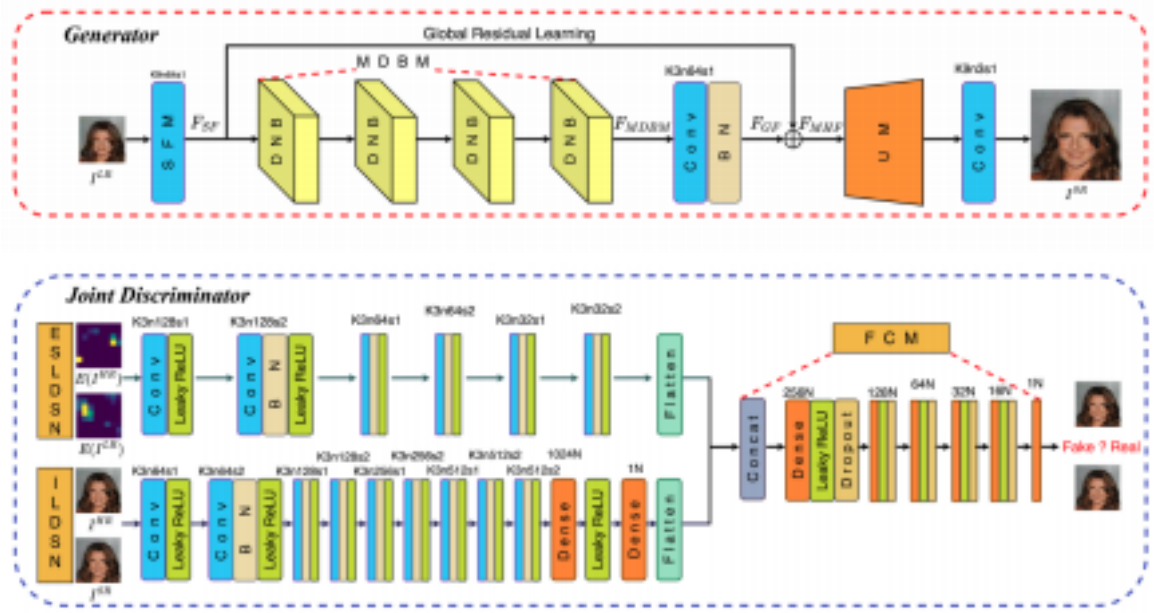

Figure 4.2: Red dotted rectangle: The architecture of Generator. Blue dotted

rectangle: The architecture of the Joint Discriminator. F_{SF} denotes shallow fea tures, F_{MDBM} denotes the outputs of MDBM, F_{GF} represents global features, and F_{MHF} represents multiple hierarchical features. K, n, and s are the kernel size, num ber of filters and strides respectively. N is the number of neurons in a dense layer.

SFM, we utilize one convolutional (Conv) layer to extract the shallow feature maps. It can be expressed as follows:

$$F_{SF} = H_{SFM}(I^{LR}) \quad (4.5)$$

where H_{SFM} represents the Conv operation in the SFM, F_{SF} denotes the shallow (low level) features, which are used for global residual learning and serve as the input to the MDBM. The following module MDBM is built up by multiple Dense Nested Blocks (DNB) formed by several RIDBs, which will be discussed in the next subsection. We introduce local residual feature extraction (LRFE) and local residual learning (LRL) in the DNB to enhance super-resolution ability and lighten training difficulty. The procedure of LRFE in i-th DNB can be formulated as:

$$F_{DNB,i\,LF} = H_{DNB,i}(H_{DNB,i-1}(\cdots(H_{DNB,1}(F_{SF}))\cdots)) \quad (4.6)$$

where $H_{DNB,i}$ acts as local residual feature extraction in the i-th DNB, which is composed of multiple blocks of RIDBs, $F_{DNB,i\,LF}$ is defined as the Local Features (LF) of the i-th DNB. Specifically, as for each DNB, it includes 3 RIDBs cascaded by residual connections and one scale layer, as shown in Figure 4.3. It can be formulated as:

$$F_{DNB,i\,\underline{L}F} = \alpha F_{i,j}(F_{i,j-1}(\cdots F_{i,1}(F_{DNB,i-1})\cdots)) \quad (4.7)$$

where $F_{i,j}$ represents the j-th RIDB of the i-th DNB. We assign α to be 0.2 in the scale layer. In order to take effective use of the local residual features, we perform the local residual learning in i-th DNB. The final output deeper features of the i-th DNB can be obtained by:

$$F_{DNB,i} = F_{DNB,i\,\underline{L}F} + F_{DNB,i-1} \quad (4.8)$$

where $F_{DNB,i}$ denotes the output deeper features of i-th DNB, which is obtained by residual connection. With the help of LRFE and LRL, the generator is able to make full use of deeper features and also efficiently propagate these features

from lower to higher layers. In our generator, there are four DNBs in MDBM, so in this case the output of 4-th DNB (i = 4) equals to the output of MDBM. Thus, the F_{MDBM} can be expressed as $F_{DNB,i=4}$. To take advantage of multi-level representations, we apply the obtained F_{MDBM} to Global Feature Fusion (GFF), where GFF is proposed to extract the global features F_{GF} by fusing feature maps produced by F_{MDBM}, the formulation is:

$$F_{GF} = H_{GF\,F}(F_{MDBM}) \quad (4.9)$$

where $H_{GF\ F}$ is a composite function of Conv layer followed by the Batch Normaliza tion (BN) layer. It aims to further extract richer features for global residual learning. Next, in order to help the generator fully use hierarchical features and alleviate gra dient vanishing problem, we adopt the Global Residual Learning (GRL) to fuse the shallow features and global features.

$$F_{MHF} = F_{SFM} + F_{GF} \quad (4.10)$$

where F_{MHF} denotes multiple hierarchical features. Next, the F_{MHF} is passed to the UM followed by one Conv layer. Then, the fused hierarchical feature is transformed from the LR space to the HR space through upsampling layers in the UM. The super-

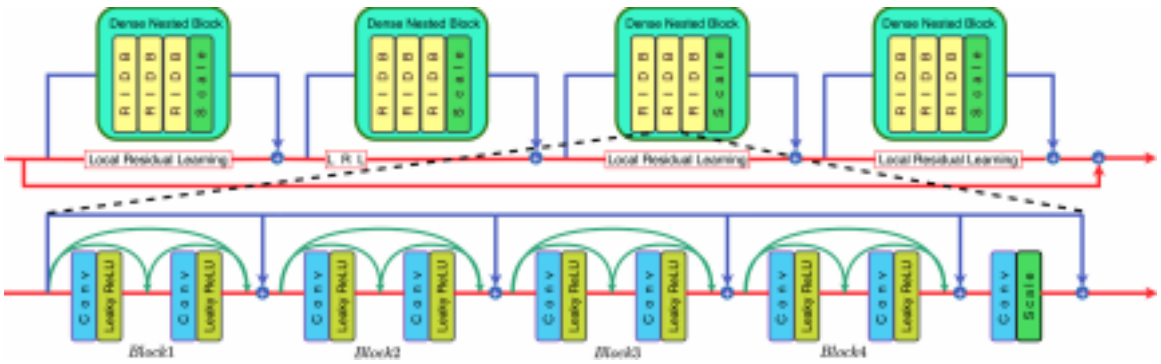

Figure 4.3: Top: Dense Nested Block (DNB) consists of multiple RIDBs. Bottom: The proposed Residual in Internal Dense Block (RIDB).

resolution process can be formulated as:

$$I^{SR} = f_{UM}(F_{MHF}) = H_{SEGA-F\,URN}(I^{LR}) \quad (4.11)$$

where f_{UM} represents the upsampling operation, $H_{SEGA-F\,URN}$ denotes the function of our method SEGA-FURN. Finally, we obtain the SR image I^{SR}.

4.4.2 Residual in Internal Dense Block

As mentioned in Section 4.4.1, the novel architecture RIDB is proposed for the gen erator, which is used to form the DNB (as shown in Figure 4.3). The proposed RIDB is able to extract hierarchical features and address the vanishing-gradient problem, which is the commonly encountered issue in [46, 72, 77, 88, 98]. The proposed RIDB is made up of four internal dense blocks and all the internal dense blocks are cas caded through residual connections performing identity mapping. The structure of the RIDB is expressed as:

$$F_{RIDB,p} = F_{p,q}(F_{p,q-1}(...F_{p,1}(F_{RIDB,p-1})...)) + F_{RIDB,p-1} \quad (4.12)$$

where $F_{RIDB,p-1}$ and $F_{RIDB,p}$ denote the input and output of the p-th RIDB respec tively, $F_{p,q}$ represents the q-th internal dense block of p-th RIDB. In addition, an internal dense block is a composition of two groups of the Conv layer followed by the LeakyReLU activation [81] layer. And the two groups are linked by dense skip

connections. Each internal dense block can be calculated as follows: $F_{q,k} =$

$$\delta(W_{q,k}[F_{q,k=1}, F_{q,k=2}]) \quad (4.13)$$

where $F_{q,k}$ represents the output of the k-th Conv layer of q-th internal dense block, $[F_{q,k=1}, F_{q,k=2}]$ refers to the concatenation of feature maps in q-th internal dense block. $W_{q,k}$ denotes the weights of the k-th Conv layer, δ denotes the LeakyReLU activation. Moreover, the residual learning and more dense connections in the RIDB effectively guarantee the feature maps of each layer are propagated into all succeeding layers, pro moting an effective way to extract hierarchical features. Thus, our proposed method is capable of obtaining abundant hierarchical feature information and alleviating the vanishing-gradient problem.

4.4.3 Semantic Encoder

The proposed semantic encoder is supposed to extract embedded semantics (as shown in Figure 4.1), which is used to project visual information (HR, LR) back to the latent space. The motivation is that the GAN-based SR models [46, 77, 88]

only exploit visual information during discriminative procedure, ignoring the semantic information reflected by latent representation. Therefore, the proposed semantic encoder will complement the missing critical property. Previous GAN's work [16, 20] has proved that the semantic representation is beneficial to the discriminator.

Based on this observation, the proposed semantic encoder is designed to inversely map the image to the embedded semantics. Significantly, the most important advan tage of the semantic encoder is that it is able to guide the discriminative process since the embedded semantics obtained from the semantic encoder can reflect semantic at tributes, such as the facial features (shape and gender) and the spatial relationship between various components of the face (eyes, mouth). It can be emphasized that the embedded semantics is fed into the joint discriminator along with HR and LR images. Thanks to this property, the semantic encoder can guide discriminator to optimize, thereby enhancing its discriminative ability.

In this context, we use two side-by-side pre-trained VGG19 [68] networks as the semantic encoder to obtain the embedded semantics of the High-Resolution (HR) face image and the Low-Resolution (LR) face image from different convolutional layers respectively. These two side-by-side VGG19 networks have the same structure except for different input dimensions, since it needs to satisfy the different dimensions of the

HR and LR face image respectively. The dimension of both two embedded semantics is 8×8×512.

4.4.4 Joint Discriminator

As shown in Figure 4.1, the proposed joint discriminator takes the tuple incorporat ing both visual information and embedded semantics as the input, where Embedded Semantics-Level Discriminative Sub-Net (ESLDSN) receives the input embedded se mantics while the image information is sent to Image-Level Discriminative Sub-Net (ILDSN). Next, through the operation of the Fully Connected Module (FCM) on a concatenated vector, the final probability is predicted. Thus, the joint discriminator has the ability to learn the joint probability distribution of image data (I^{HR}, I^{SR}) and embedded semantics ($E(I^{HR})$, $E(I^{LR})$). There are two sets of paths entering into the joint discriminator. The set of paths

shown in red indicates a real tuple which consists of a real sample I^{HR} from the dataset and its embedded semantics $E(I^{HR})$. For the blue path, a fake tuple is constructed from SR image I^{SR} generated from generator and $E(I^{LR})$ obtained from LR image through semantic encoder. As a result, different from [13, 46, 77, 88], our joint discriminator has the ability to evaluate the difference between real tuple (I^{HR}, $E(I^{HR})$) and fake tuple (I^{SR}, $E(I^{LR})$).

Moreover, in order to alleviate the problem of gradient vanishing and enhance the model stability, we adopt the Relativistic average Least Squares GAN (RaLSGAN) objective loss for the joint discriminator by applying the RaD to the least squares loss function [55]. Let's denote the real tuple by $X_{real} = (I^{HR}, E(I^{HR}))$ and denote the fake tuple by $X_{fake} = (I^{SR}, E(I^{LR}))$. The process that makes joint discriminator to be relativistic can be expressed as follows:

$$\tilde{C}(X_{real}) = (C(X_{real}) - E_{xf}[C(X_{fake})])$$
$$\tilde{C}(X_{fake}) = (C(X_{fake}) - E_{xr}[C(X_{real})]) \quad (4.14)$$

where $\tilde{C}(\cdot)$ denotes the probability predicted by joint discriminator, E_{xf} and E_{xr} describe the average of the SR images (fake) and HR images (real) in a training batch. Moreover, the least squares loss is used to measure the distance between HR and SR images. According to Equation 4.15, we optimize the joint discriminator by

adversarial loss L^{RaLS}_{D} and the generator is updated by L^{RaLS}_{G}, as in Equation 4.16.

$$L^{RaLS}_{D} = E_{I^{HR}\sim p(I^{HR})}[(\tilde{C}(X_{real}) - 1)^2] + E_{I^{SR}\sim p(I^{SR})}[(\tilde{C}(X_{fake}) + 1)^2] \quad (4.15)$$

$$L^{RaLS}_{G} = E_{I^{SR}\sim p(I^{SR})}[(\tilde{C}(X_{fake}) - 1)^2] + E_{I^{HR}\sim p(I^{HR})}[(\tilde{C}(X_{real}) + 1)^2] \quad (4.16)$$

where $I_{HR} \sim P_{I^{HR}}$ and $I_{SR} \sim P_{I^{SR}}$ indicate the HR images and SR images distribution respectively. Furthermore, with the help of least squares loss and relativism in

RaLS, SEGA-FURN is remarkably more stable and generates authentic and visually pleasant SR images.

The joint discriminator consists of two sub-nets, Embedded Semantics-Level Dis criminative Sub-Net (ESLDSN), Image-Level Discriminative Sub-Net (ILDSN) and one Fully Connected Module (FCM). The ESLDSN takes the embedded semantics as the input and downsamples it through 6 convolutional layers with 3x3 kernels and the stride of 1 or 2 alternately, and then reshapes it to a 32-dimensional vector. The ILDSN receives an image, and performs feature extraction through 9 groups of con volutional layers using the kernel of the size 3x3 followed by the LeakyReLU [81] and the Batch Normalization (BN) layer to obtain the flattened 64-dimensional vector. Next, the resulting two vectors are concatenated by the concatenation layer, and then fed into the FCM. As for FCM, it contains six dense layer blocks, where each dense layer block includes a dense layer, a LeakyReLU activation layer and a dropout layer except for the last single dense layer. These six dense layers have 256, 128, 64, 32, 16 and 1 neurons respectively. Finally, the output is a probability that how the given HR face image is more realistic than the SR face image.

4.4.5 Feature Extractor

We further exploit pre-trained VGG19 [68] network as feature extractor φ in SEGA FURN to obtain feature representations used to calculate the perceptual loss $L_{perceptual}$, where $L_{perceptual}$ is utilized in SEGA-FURN to eliminate the facial ambiguity and re cover missing details of SR images. It is measured as the Euclidean distance between two feature representations of SR images and HR images. Instead of using high-level features as in SRGAN, ESRGAN for perceptual loss, we adopt low-level features be fore activation layer (i.e., feature representations from ‘Conv3 3’ layer in the feature

extractor), which contains complex edge textures.

4.4.6 Loss Function

We involve perceptual loss $L_{perceptual}$ to constrain the intensity and feature similarities between HR and SR images [19, 38]. Furthermore, adversarial loss L^{RaLS}_{G} is adopted

to super-resolve SR images containing visually appealing details and faithful to the HR images.

perceptual Loss: $L_{perceptual}$ is able to reduce the gap between SR image and HR image. It is formulated as:

$$L_{perceptual} = \frac{1}{WH} \sum_{q=1}^{W} \sum_{r=1}^{H} (\varphi_{i,j}(I^{HR})_{q,r} - \varphi_{i,j}(I^{SR})_{q,r})^2 \quad (4.17)$$

where W, H describe the height and width of the feature maps, $\varphi(\cdot)$ denotes the output of feature extractor, $\varphi_{i,j}$ indicates the feature representations obtained from j-th convolution layer before i-th maxpooling layer.

Total Loss: The total loss function L_{total} for the generator can be represented as a weighted combination of two parts: perceptual loss $L_{perceptual}$ and adversarial loss L^{RaLS}_{G}, the formula is described as follows:

$$L_{total} = \lambda_{per} L_{perceptual} + \lambda_{adv} L^{RaLS}_{G} \quad (4.18)$$

where λ_{per}, λ_{adv} are the trade-off weights for the $L_{perceptual}$ and the L^{RaLS}_{G}. We set λ_{per}, λ_{adv} empirically to 1 and 10^{-3} respectively.

4.5 Experiments

In this section, we first present the details of dataset and training implementation. Then, we demonstrate the experiments and evaluation results. We further compare our method with state-of-the-art methods. Moreover, in order to prove the effective ness of SEGA-FURN, we conduct ablation experiments to verify the contributions of the components proposed in this work.

We compare the proposed method with the state-of-the-art SR methods [1, 8, 17, 42, 46, 50, 53, 73, 77, 82, 84, 86, 87, 88, 89, 90, 102] quantitatively and qualitatively. Please note that we involved some experimental results which were demonstrated in

their corresponding published papers. In addition, the quantitative results of some state-of-the-art methods were missing when upscaling factor is 4×, because they

did not conduct experiments for upscaling factor 4×, and some methods did not exploit Structural Similarity (SSIM) as the evaluation criterion. For the qualitative com parison, we adopted the traditional super-resolution method, Bicubic interpolation, and the Generative Adversarial Network (GAN) [26] -based super-resolution method, SRGAN [46] and ESRGAN [77]. In order to compare the visual effects fairly, we used their published code of the above models.

4.5.1 Datasets

We conducted experiments on the public large-scale CelebFaces Attributes dataset, CelebA [52]. It consists of 200K celebrity face images of 10,177 celebrities. We used a total of 202,599 images, where we randomly selected 162,048 HR face images as the training set, and all the rest 40,511 images were chosen as the testing set.

4.5.2 Implementation Details

To verify the effectiveness of our method, we conducted experiments with multiple upscaling factors 4× and 8× respectively. We resized and cropped the images to 256x256 pixels as our HR face images without any alignment operation. In order to obtain two groups of LR downsampled face images, we used a bicubic interpolation method with downsampling factor r =4 to produce 64x64 pixels, and factor r=8 to produce 32x32 pixels LR images.

We trained our network 30k epochs using the Adam optimizer [44] by setting β_1 =0.9, β_2=0.999 with the learning rate of 10^{-4} and batch size of 8. We alternately updated the generator and discriminator until the model converged. For the quan titative comparison, we adopted Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) as the evaluation metrics.

4.5.3 Qualitative Comparison

The 4× and 8× qualitative results are depicted in Figure 4.4 and 4.5 respectively. The Figure 4.4 shows the 4× visual results. As for Bicubic interpolation, we observe that its results contain over-smooth visualization effects. SRGAN [46] relatively enhances the SR results compared to Bicubic interpolation, but it still fails to generate fine

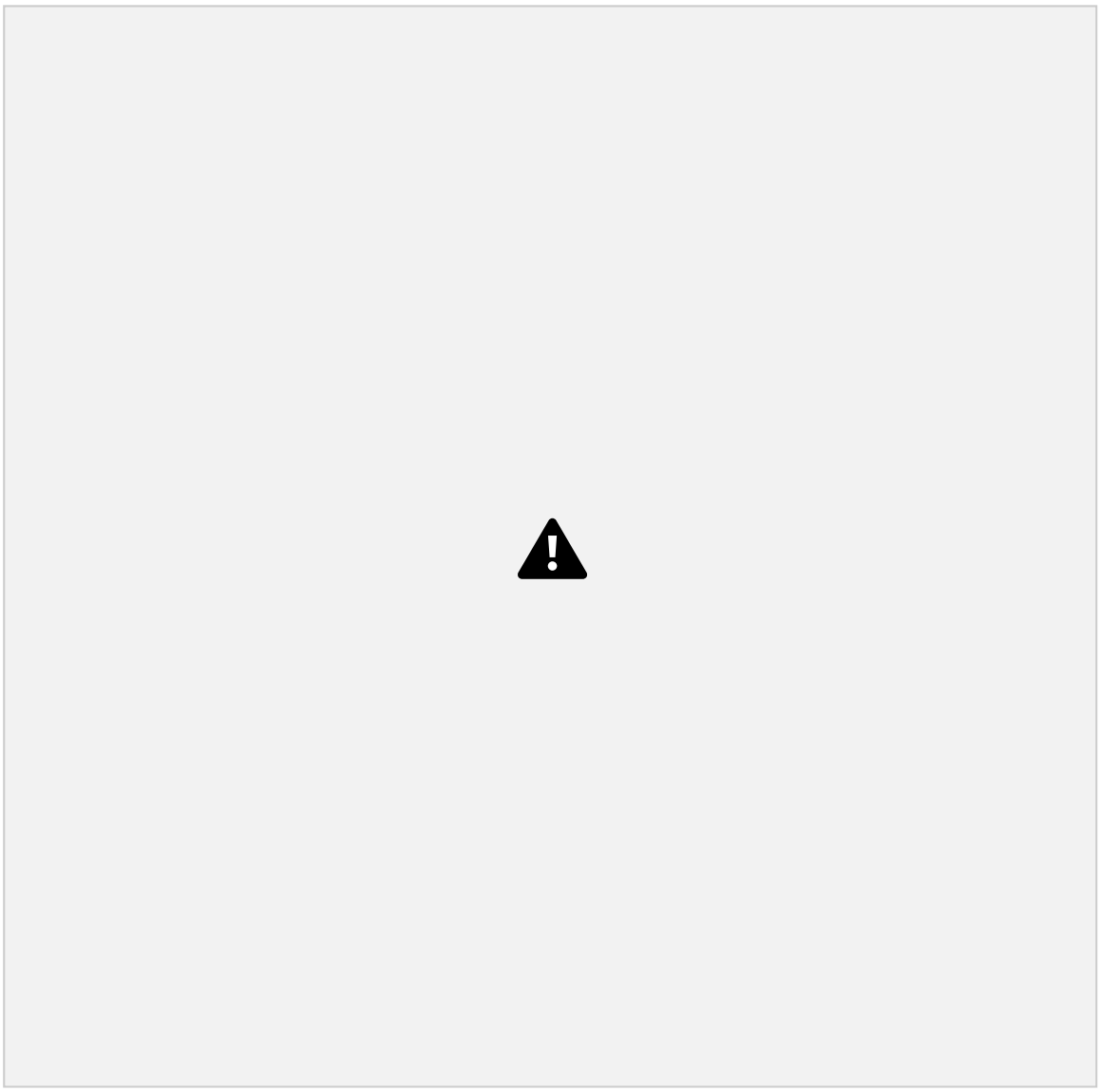

Figure 4.4: Qualitative comparison against state-of-the-art methods. The results of 4× upsampling factor from 16x16 pixels to 256x256 pixels. From left to right: (a) HR images, (b) LR inputs, (c) Bicubic interpolation, (d) Results of SRGAN [46], (e) Results of ESRGAN [77], and (f) Our method.

details especially in facial components, such as eyes, and mouth. It is obvious that ESRGAN [77] produces overly smoothed visual results and misses specific textures. On the contrary, the 4× SR images produced by our method retain facial-specific details and are faithful to HR counterparts.

To reveal the powerful super-resolution ability of our proposed method, we further

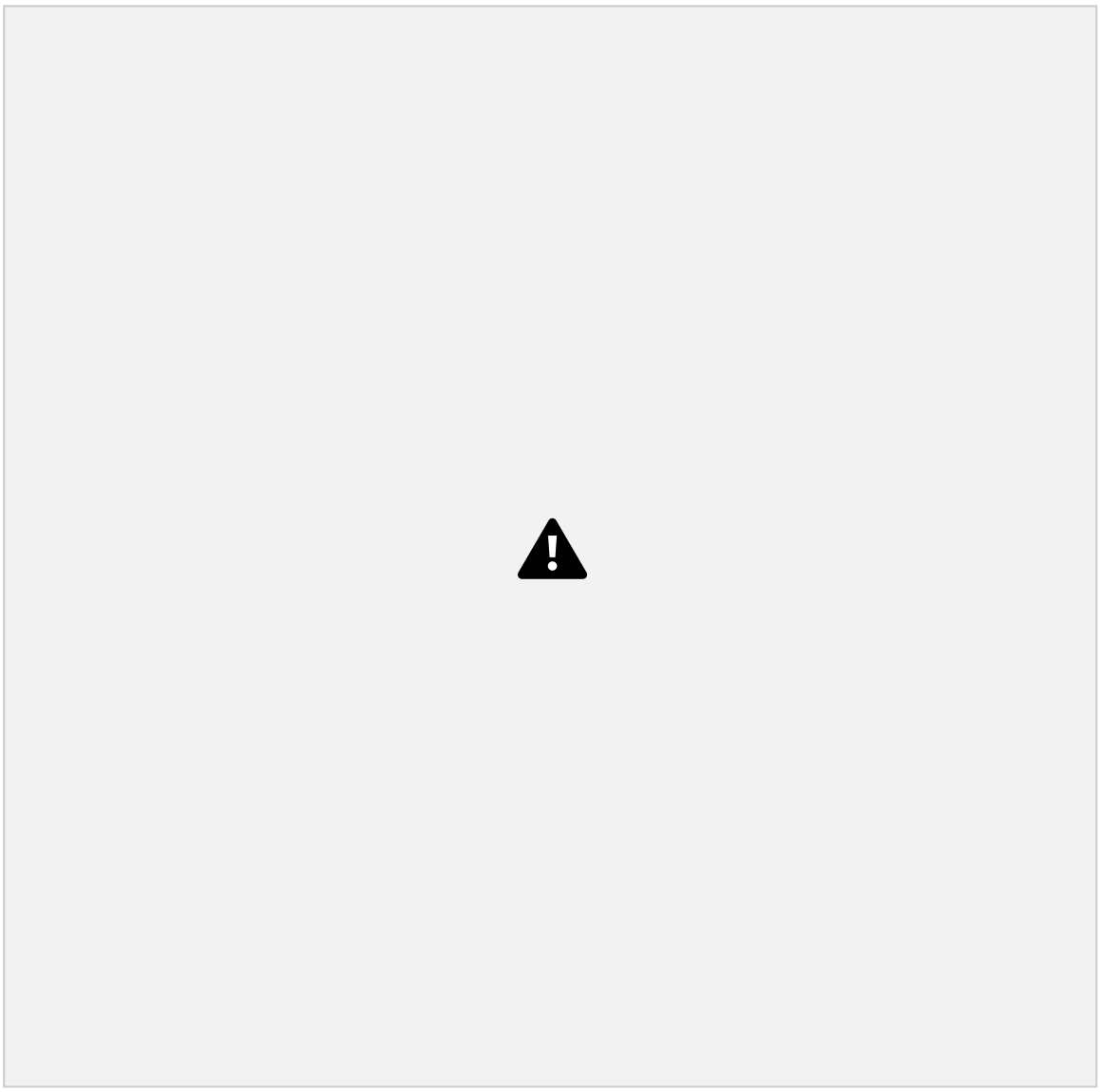

Figure 4.5: Qualitative comparison against state-of-the-art methods. The results of 8× upsampling factor from 16x16 pixels to 256x256 pixels. From left to right: (a) HR images, (b) LR inputs, (c) Bicubic interpolation, (d) Results of SRGAN [46], (e) Results of ESRGAN [77], and (f) Our method.

conduct experiments with 8× ultra upscaling factor. As shown in the Figure 4.5, it apparently presents that the SR visual quality obtained by Bicubic interpolation, SRGAN and ESRGAN is decreased, since the magnification is increased, resulting in the correspondence between HR and LR images incompatible. The outputs of Bicubic interpolation generate unpleasant noises. SRGAN encounters a mode collapse problem during the super-resolution process, so that it produces severe distortions in

Method	CelebA 4×		CelebA 8×	
	PSNR	SSIM	PSNR	SSIM
Bicubic	26.50	0.79	22.90	0.65
Yang et al. [84]	-	-	21.35	0.60
SRCNN [17]	28.93	0.79	23.11	0.65
Ma et al. [53]	-	-	23.12	0.64
CSGM-BP [8]	26.44	-	22.71	-
DIP [73]	27.35	-	23.45	-
Yang et al. [82]	-	-	23.07	0.65
CBN [102]	29.37	0.79	18.77	0.54
URDGN [88]	29.10	0.79	24.82	0.70
IAGAN [1]	27.16	-	23.49	-
TDN [89]	-	-	22.66	0.66
FSRFCH [86]	-	-	23.14	0.82
SRGAN [46]	26.76	0.82	20.64	0.62
ESRGAN [77]	23.82	0.71	20.32	0.57
TDAE [90]	29.49	0.81	20.40	0.57
FaceAttr [87]	29.78	0.82	21.82	0.62
WGAN-GP [14]	23.28	0.69	-	-
VDSR [42]	-	-	19.58	0.57
Liu et al. [50]	-	-	21.60	0.66
Ours	30.14	0.87	25.21	0.73

Table 4.1: Quantitative comparison on CelebA dataset for upscaling factor 4× and 8×, in terms of average PSNR(dB) and SSIM. Numbers in bold are the best evaluation results among state-of-the-art methods.

SR images. As for ESRGAN, it produces the SR images which show broken textures, noticeable artifacts around facial components. In contrast, our method is capable of producing photo-realistic SR images which preserve perceptually sharper edges and fine facial textures.

4.5.4 Quantitative Comparison

The quantitative results with multiple ultra upscaling factors 4× and 8× are shown in Table 4.1. It is obvious that our method attains the best in both PSNR and SSIM evaluations, 30.14dB/0.87 for 4× and 25.21dB/0.73 for 8×, among all other methods. FaceAttr [87] is the second best method for 4×, 29.78dB/0.82, however, it degrades dramatically and performs poorly when the upscaling factor increases to 8×, obtaining 21.82dB/0.62. In contrast, our proposed method ranks the first for both upscaling factors 4× and 8×, which reflects the robustness of the proposed

Variant	RIDB	SE	RaLS
(A) RIDB-Net	√		
(B) RIDB-RaLS-Net	√		√
(C) RIDB-SE-Net	√	√	
(D) RIDB-SE-RaLS-Net (SEGA-FURN)	√	√	√

Table 4.2: Description of SEGA-FURN variants with different components in exper iments.

SEGA-FURN. Moreover, it is notable that our proposed method not only boosts PSNR/SSIM by a large margin of 1.04dB/0.08 over the classic method URDGN [88] with upscaling factor 4× but also is higher than URDGN which is the second best for the 8× upscaling. This observation shows a stable ability of our method with multiple upscaling factors. In addition, we compare with SRGAN [46] and ESRGAN [77] which also uses generative adversarial structure. It is obvious that our method not only improves SR image quality from perceptual aspects but also achieves impressive numerical results.

4.6 Ablation Study

We further implemented ablation studies to investigate the performance of the pro posed method. As shown in Table 4.2, we list several variants based on different proposed components. First, among them, RIDB-Net can be used as the baseline variant, which only contains a single component RIDB. Second, the RIDB-RaLS-Net is constructed by removing the Semantic Encoder (SE) from the SEGA-FURN. Next, RIDB-SE-Net means to remove RaLS loss of SEGA-FURN, and RIDB-SE-RaLS-Net equals to SEGA-FURN including all the three components. In addition, we provide the upscaling factor 4× and 8× visual results of these variants in Figure 4.6 and 4.7 respectively, and quantitative comparison in the Table 4.3.

4.6.1 Effect of RIDB

We compare the proposed RIDB with other feature extraction blocks, such as Residual Block (RB) from SRGAN [46] and Residual in Residual Dense Block (RRDB) of ESRGAN [77]. As shown in Figure 4.4 and Table 4.1, it is noticeable that the SR results generated by our generator employing RIDB outperforms SRGAN utilizing RB and ESRGAN adopting RRDB both in qualitative and quantitative comparisons. The

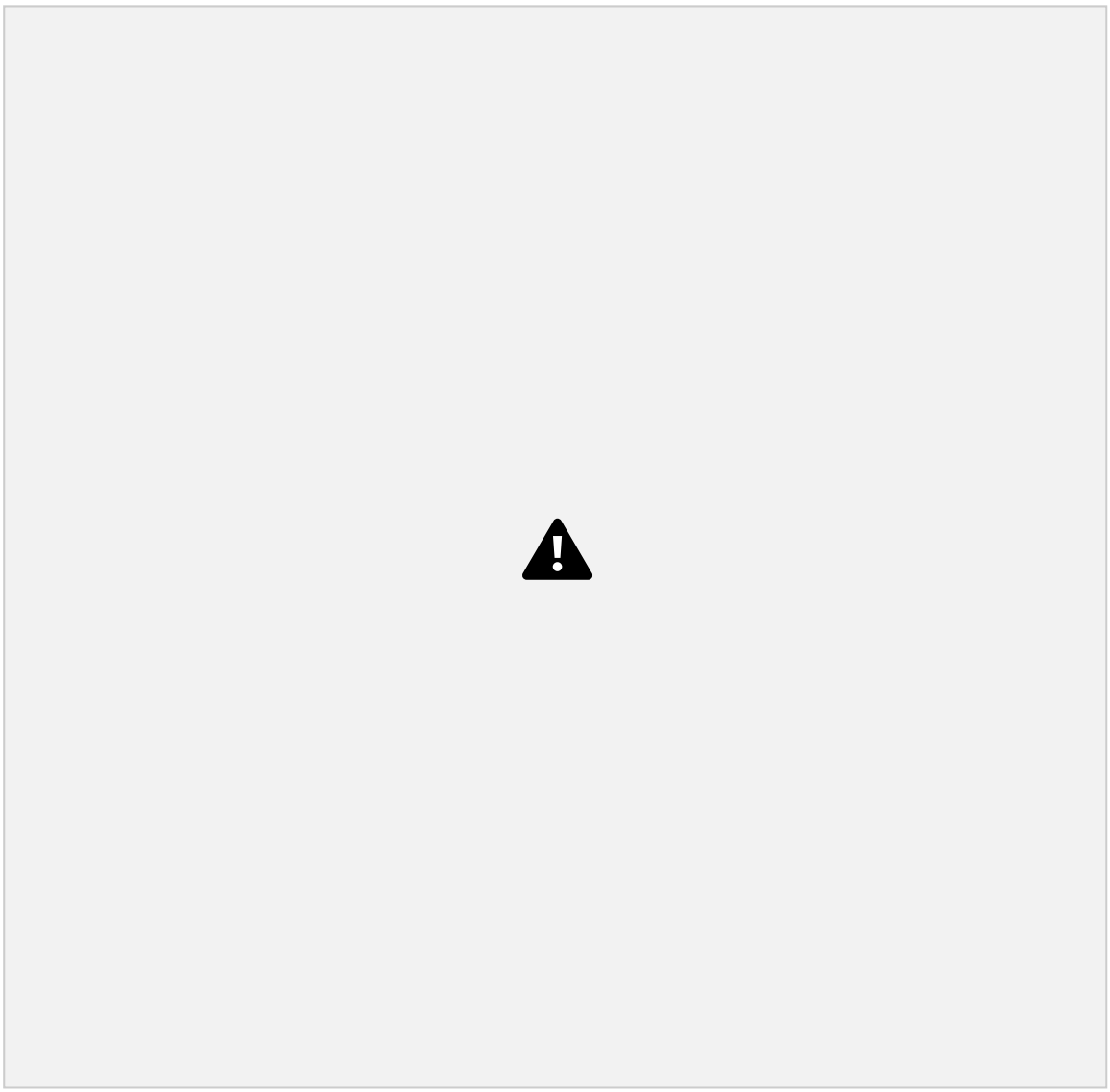

Figure 4.6: Qualitative comparison of ablation studies. The results of upscaling factor 4×. From left to right: (a) HR images, (b) LR inputs, (c) Results of RIDB-Net, (d) Results of RIDB-RaLS-Net (e) Results of RIDB-SE-Net, and (f) Results of RIDB SE-RaLS-Net (SEGA-FURN).

reason is that our RIDB introduces densely connected structure to combine different level features, but there are no dense connections in RB. In addition, different from RRDB, the proposed RIDB designs multi-level residual learning within each basic internal dense block, which is able to boost the flow of features through the generator and provide hierarchical features for the super-resolution process. Based on these observations and investigations, it is persuasive to validate the effectiveness of the

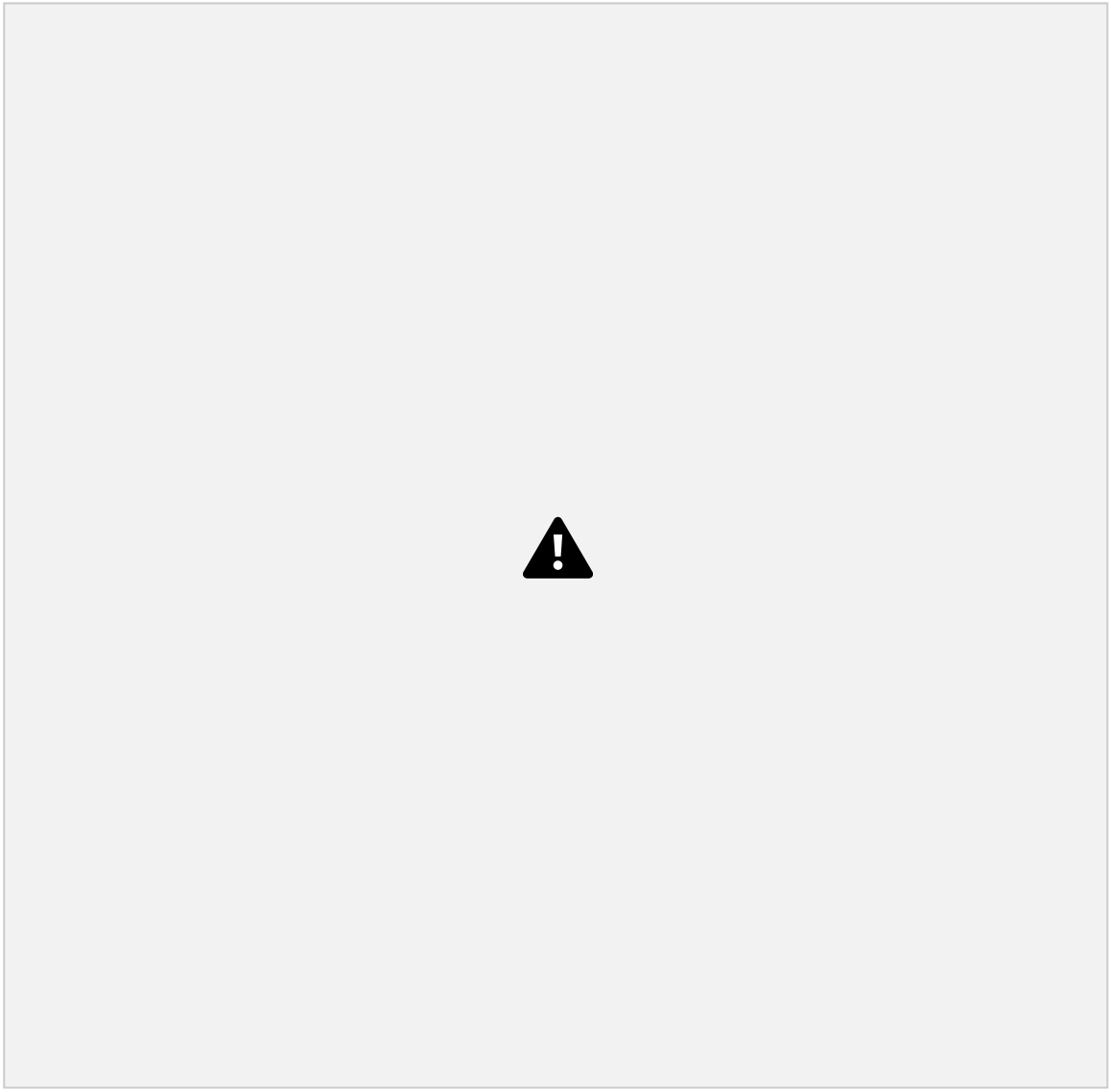

Figure 4.7: Qualitative comparison of ablation studies. The results of upscaling factor 8×. From left to right: (a) HR images, (b) LR inputs, (c) Results of RIDB-Net, (d) Results of RIDB-RaLS-Net (e) Results of RIDB-SE-Net, and (f) Results of RIDB SE-RaLS-Net (SEGA-FURN).

proposed RIDB.

4.6.2 Effect of SE

The Ablation (A) and (C) performed by RIDB-Net and RIDB-SE-Net aim to illus trate the advantage of SE and also verify the effectiveness of the joint discriminator. The RIDB-SE-Net can obtain embedded semantics extracted by

SE and further feed

51

Ablation	CelebA 4×	CelebA 8×
	PSNR/SSIM	PSNR/SSIM
(A) RIDB-Net	28.64/0.8514	24.25/0.7177
(B) RIDB-RaLS-Net	28.71/0.8526	24.37/0.7218
(C) RIDB-SE-Net	29.60/0.8607	24.44/0.7181
(D) RIDB-SE-RaLS-Net	30.14/0.8682	25.21/0.7250

Table 4.3: Quantitative comparison of different variants on CelebA dataset for up scaling factor 4× and 8×.

these semantics along with image data to the joint discriminator. In training pro cess, the embedded semantics is capable of providing useful semantic information for the joint discriminator. Such innovation can enhance the discriminative ability of the joint discriminator. Compared with RIDB-Net which does not employ SE and joint discriminator, the RIDB-SE-Net achieves significant improvements in terms of quantitative comparisons. Furthermore, as shown in Figure 4.6 and 4.7 (4× and 8× visual results), there is also a noticeable refinement in detailed texture. The enhanced performance can verify that the extracted embedded semantics has superior impact on SR results and the SE along with the joint discriminator play a critical role of the proposed method.

4.6.3 Effect of RaLS

The ablation (A) and (B) are conducted to demonstrate the effect of RaLS loss. We replace the RaLS loss of RIDB-Net with the generic GAN loss, Binary Cross Entropy (BCE) and keep all the other components the same. As shown in Table 4.3, it is obvious that once we remove the RaLS loss in RIDB-Net, the quantitative results are lower than RIDB-RaLS-Net which has RaLS loss. As expected, BCE used in RIDB Net shows unrefined textures. In contrast, when RaLS is utilized in variant, the visual results are perceptually pleasing with more natural textures and edges. Thus, it can demonstrate that the RaLS loss is capable of greatly improving the performance of super-resolution.

4.6.4 Final Effect

From the comparison between ablation (D) and other studies, it is obvious that

the large enhancement is noticeable by integrating all these three components. Finally,

we refer the RIDB-SE-RaLS-Net to SEGA-FURN which is the ultimate proposed method.

4.7 Conclusions

In this chapter, we proposed a novel Semantic Encoder guided Generative Adversarial Face Ultra-resolution Network (SEGA-FURN) to super-resolve a tiny LR unaligned face image to its HR version with multiple large ultra-upscaling factors (e.g., 4× and 8×). Owing to the proposed Semantic Encoder, Residual in Internal Dense Block and the Joint Discriminator adopting RaLS loss, our method successfully produced photo-realistic SR face images. Extensive experiments and analysis demonstrated that SEGA-FURN is superior to the state-of-the-art methods.

Chapter 5

Real-World Image Super Resolution via Unsupervised Bi-directional Cycle Domain Transfer Learning based

Generative Adversarial Network

5.1 Overview

In this chapter, we concentrate on real-world natural image super-resolution in an un supervised manner. Deep Convolutional Neural Networks have exhibited impressive performance on image super-resolution by reconstructing a high resolution image from a low resolution image. However, most state-of-the-art methods heavily rely on two limited properties where the training LR and HR images are paired and artificially pre-determined by known degradation kernel (e.g., bicubic downsampling) to train

the networks in the fully supervised fashion. As a result, existing methods fail to deal with real-world super-resolution tasks, since the paired LR and HR images in real world are typically unavailable and degraded by the complicated and unknown ker nel(s). To break these restrictions, in this chapter, we propose the Unsupervised Bi directional Cycle Domain Transfer Learning-based Generative Adversarial Network (UBCDT-GAN), which has the ability to super-resolve HR image from the real-world LR image with complex and inevitable sensor noise in an unsupervised manner. Our proposed method consists of an Unsupervised Bi-directional Cycle Domain Trans fer Network (UBCDTN) and Semantic Encoder guided Super Resolution Network (SESRN). Firstly, the UBCDTN is able to produce an approximated real-like LR image through transferring the LR image from an artificially degraded domain to the real-world LR image domain with natural characteristics. Secondly, the SESRN takes the approximated real-like LR image as input and super-resolves it to a photo-realistic HR image. Extensive experiments on unpaired real-world image benchmark datasets demonstrate that the proposed method achieves promising performance compared to state-of-the-art methods.

5.2 Introduction

Single Image Super-Resolution (SISR) aims to reconstruct a High-Resolution (HR) image from a single Low-Resolution (LR) image version, which has been a prosperous research topic in recent years. It has been widely applied in many computer vision applications, such as surveillance [62], image enhancement [11] and medical image processing [66]. In the SISR task, the general degradation formula is expressed as:

$$y = (x \overset{O}{} k) \downarrow s + n \quad (5.1)$$

where x represents the HR image, y is the degraded LR image, k denotes a blur kernel, and $\overset{N}{}$ is the convolution operation performed on x and k. $\downarrow s$ denotes a downsampling operation of the image with scale factor s, and n is considered as an additive white Gaussian noise. However, under the real-world scene setting, n should take into account many possible conditions such as sensor noise, compression artifacts, and unpredicted noise caused by physical devices. Resulting from the existence of uncertain noise n, the SISR has become a particularly ill-posed inverse task since there are infinite HR images that can be recovered from a given LR image, in which

it is required to select the most plausible solutions.

With the extraordinary development of deep learning techniques, a great number of deep learning-based SISR models have been proposed, such as RDN [98], EDSR [48], SRDenseNet [72], SRGAN [46], ESRGAN [77]. Despite the successful progress achieved by the aforementioned methods, there still exists unnoticed issues which should be considered. All the aforementioned SR methods trained on supervised manner with a large number of paired images, synthesized LR images and its HR version, resulting in deteriorative performance when they are applied to real-world scenarios. The limitation is that the paired training data is unavailable and the degradation of the input LR image is unknown in the real-world scenes.

Owing to the absence of the real-world LR and its HR counterpart paired data, the super-resolution fidelity of training and evaluation results is impeded. In the past few years, several methods tried to collect real-world LR and HR paired datasets to improve the poor generalization when dealing with unsupervised super-resolution learning. Cai *et al*. [10] first collected the RealSR dataset

comprising paired HR and LR data manually and then proposed a Laplacian pyramid based kernel prediction network to recover the HR images. Zhang *et al*. [96] utilized super-resolution raw data to produce a dataset SR-RAW consisting of optical super-resolution ground truth. However, unfortunately, it is noted that it is difficult and impracticable to make a pair between two different domains, e.g, HR and LR image data, under real world scenarios, such as medical images and satellite images.

Due to the lack of real-world LR and HR paired data, numerous supervised meth ods were trained on artificially synthesized image pairs. Trained with a large number of image pairs generated by pre-determined degradation operation (e.g., bicubic) on the HR images, the proposed SRDenseNet [72], RDN [98] and LapSRN [45] can learn richer feature representations and recover more image details. The problem is that by imposing a pre-determined degradation operation (e.g, bicubic downscaling) on the HR images, the degraded LR training images can be obtained. However, it is unreasonable to simply apply the input images downsampled by the ideally fixed bicubic kernel to the training and testing phase [41, 54]. There still exists a large domain gap between real-world LR images and artificial LR images downscaled by pre-defined degradation kernel. In addition, the artificial LR images may eliminate diverse patterns and complicated characteristics belonging to real-world LR images such as sensor noise, unpredicted artifacts, and natural characteristics. Overall, the existing SR methods normally encounter a serious domain consistency problem and

produce a poor performance in practical scenarios [92, 99].

Recently, instead of utilizing any known degradations during data processing, many unsupervised SR methods have been trained on the real-world datasets [32, 56, 67, 75], which aims to improve the SR performance in such real-world applications. Yu *et al.* [32] presents the soft maximum a posteriori (MAP) estimation framework to estimate the reasonable blur kernel, in which it can perform blur identification and HR reconstruction. The iterative kernel estimation algorithm proposed by Tomer *et al.* [56] can recover SR blur kernel by inputting LR image directly, which can improve reconstruction performance compared with previous methods utilized by certain blur kernels. However, Since it is difficult to estimate downscaling kernels precisely, these methods still fail to deal with real-world noise expectably.

Therefore, it is imperative to explore an effective method which can apply un paired images to satisfy the need for real-world SR scenarios. It must be different from the aforementioned SR methods which do not take into consideration the do main bridge between the LR images generated from a known degradation (e.g., bicu bic downscaling) and the LR images from real-world. To address the above limita tions, in this chapter, we proposed Unsupervised Bi-directional Cycle Domain transfer Learning-based Generative Adversarial Network (UBCDT-GAN) for real-world im age super-resolution. The proposed method consists of two networks, Unsupervised Bi-directional Cycle Domain Transfer Network (UBCDTN) and Semantic Encoder guided Super Resolution Network (SESRN). To simulate the real-world data distri bution and reduce the domain gap between the generated LR image domain and the real world LR image domain, we first designed UBCDTN to estimate the inherent degradation kernel from real-world LR distribution and translate the artificially de graded LR domain image to the real-world domain. With the help of cycle consistency mechanism [101], the proposed UBCDTN is able to learn bi-directional inverse map ping in an unsupervised manner, which can ensure the generated real-like LR images preserves desired characteristics of real-world patterns. Besides, we also enforced aux iliary constraints on the UBCDTN such as adversarial loss, identity loss, and percep tual loss. The designed domain transfer network provides an effective way to generate real-like LR images, which can construct the paired real-world LR-HR data for the followed SESRN. For the second step, we employed the previous proposed Seman tic Encoder guided Generative Adversarial Face Ultra-resolve Hallucination Network (SEGA-FURN) as the Super Resolution Network, namely Semantic Encoder guided Super Resolution Network (SESRN) in this case. The goal of the SESRN is to super-

resolve the real-like LR images to the photo-realistic HR images. We evaluated our proposed method on the NTIRE 2020 Super Resolution Challenge Track 1 validation dataset and the quantitative and qualitative comparisons demonstrate the superior ity of the proposed UBCDT-GAN compared with other state-of-the-art methods. A comparison of visual results with various latest methods is shown in Figure 5.9 and 5.10, and the numeric results can be seen in Table 5.1 and 5.2.

The main contributions of our proposed method can be summarized as

follows: 1) We proposed a novel bi-directional cycle domain transfer network, UBCDTN. According to the domain transfer learning scheme, the designed bi-directional cycle architecture is able to eliminate the domain gap between the generated real-like LR images and real-world images in an unsupervised manner.

2) We further imposed the auxiliary constraints on the UBCDTN by incorporating adversarial loss, identity loss, and perceptual loss, which can guarantee that the real like LR images contain the same style as real-world images.

3) We adopted the previous proposed SESRN as a deep super-resolution network to generate visually pleasant SR results under the supervised learning settings. 4) Benefiting from the collaborative training strategy, the proposed UBCDT-GAN is able to train in an end-to-end fashion, which is able to ease the entire training procedure and strengthening the robustness of the model.

5.3 Related Work

In this section, we present a brief literature review of previous works related to our proposed method. We present state-of-the-art deep learning based supervised meth ods for SISR. Then, we introduce the latest unsupervised learning methods for real world scenes. In addition, since our method also deals with image domain translation, we further present the typical image-to-image translation methods.

5.3.1 Paired Image Super Resolution

In recent years, deep learning-based methods have exhibited exceptional popularity and extraordinary ability to enhance SISR performances. Most of these methods rely on supervised settings, where the specific pre-defined and paired LR and HR images are obtained in the training and testing dataset. Thus, methods can learn the mapping between LR and HR paired images in the training process. There are

many typical methods such as SRCNN [17], VDSR [42], DRRN [70], EDSR [48]. However these methods heavily rely on L1 or L2 losses to optimize models alone, producing unexpected results. To address this limitation, many GAN-based methods are proposed, such as SRGAN [46], ESRGAN [77] and URDGN [88].

The details of these methods can be seen in chapter 1.

5.3.2 Blind Image Super Resolution

Although the aforementioned methods achieved noticeable progress in the SR field, they have the limited ability to solve blind SISR problems. Specifically, the blind SISR is defined as the task that supposing paired LR and HR data is processed by unknown degradation and downsampling kernel. By contrast, all the above methods are specifically trained and tested on synthesized datasets where the paired LR-HR images are performed through simple known degradation (e.g., Bicubic, Linear). Since these methods have never seen practical artifacts and characteristics, they cannot apply to real world scenes where the image data contains natural noise and diverse degradation types. Thus, many researchers attempt to solve blind SISR problems where the degradation kernel executing on LR images is unavailable. The Iterative Kernel Correction method (IKC) [27] was proposed by Gu *et al*. to estimate blur kernel and eliminate artifacts caused by kernel discrepancy. Based on the previous SR results, the estimated kernel is corrected until the approximated blur kernel similar to real ones, leading to removing normal artifacts. Lugmayr *et al*. [101] further proposed CycleGAN, which is able to simulate the practical degradation kernel and transform the HR images to the LR images containing real world characteristics. Zhang *et al*. [94] proposed IRCNN which includes a set of CNN denoisers. To estimate the blur model, the IRCNN incorporates a learned denoiser into the model-based optimization method. However, the results of IRCNN indicate that it is difficult to estimate a comprehensive degradation kernel in real world conditions. The blind SR methods have the limited capability to approximate the unknown degradations. Thus, there is still room for improvement in dealing with blind image super-resolution.

5.3.3 Unpaired Super Resolution

As mentioned above, all the aforementioned deep learning-based SR methods are trained in the supervised fusion in which the paired LR-HR images are required and LR images are simply generated by bicubic downsampling. Since these supervised

methods require massive paired and synthesized LR-HR data during training, they perform poorly when dealing with real world SR problems.

To solve this task, one solution is to create paired real world LR-HR data. Cai *et al*. [9] adjusted the focal length of a digital camera to capture paired LR-HR images at the same scene, creating a real world dataset. However, collecting a comprehen sive real LR-HR paired datasets with diverse degradations is expensive and impossi ble. Moreover, these methods heavily rely on large-scale data collection mechanisms, which require complicated hardware, being incompatible with real-world needs. Thus, recently, many unsupervised methods are proposed to satisfy real world conditions where the paired LR-HR image data is unavailable and the input image is corrupted by unknown degradation kernels. In [67], Shocheret *et al*. proposed a zero-shot super-resolution method (ZSSR) which is an image-specific method. By employing the internal recurrence of information inside an image, ZSSR can use pseudo image pairs to recover LR images with diverse blur kernels. However, since the inference process of ZSSR is in real time, which greatly increases inference time, ZSSR may not fulfill the needs of real scenes. Inspired by CycleGAN, Yuan *et al*. [92] proposed the method which included two sets of generator and discriminator, namely CinCGAN. This method first generated bicubic downsampled LR images from the input and then super-resolved LR images to HR images. However, the CinCGAN only considered single bicubic degradation, resulting in poor generalization in complicated real world SR tasks. By contrast, our method utilizes UBCDTN to simulate diverse real world degradation types, making it perform well in the real world. Fritsche *et al*. [23] proposed DSGAN, which incorporated the frequency separation network into GAN to recover high frequencies of SR results. In [63], Ren *et al*. first created an un paired mobile SR dataset from registered mobile-DSLR images and then fine-tuned a generic SR model on this created dataset, which can improve the visual quality of mobile images. However, since these unsupervised methods merely concentrate on specific real world SR tasks, such as bicubic images and mobile LR images, it still has the limited generalization ability to handle severely degraded LR images containing diverse degradation kernels. On the contrary, in our method, we design the UBCDTN to approximate comprehensive real world degradation kernels and produce translated real-like LR images containing sensor noise and natural artifacts which are the same as real world LR images. Then the proposed SESRN is able to

super-resolve high quality HR images from the real-like LR version.

5.3.4 Image-to-Image Translation

The image super-resolution task can be considered as the specific image translation problem, where both tasks aim to translate the image from the source domain to the target domain. As for image super-resolution, it attempts to translate images from source LR domain to target HR domain. However, super-resolution problems are more challenging than the Image-to-Image translation. Since the problem of the Image-to-Image translation receives the input image and produces the output image with the same size of the input image, it mainly concentrates on style translation. However, the super-resolution problem not only produces the output image which is several times larger than input images but also produces photo-realistic SR images with accurate style and natural textures.

There are some typical Image-to-Image translation methods. Isola *et al*. [37] proposed the pix2pix GAN to translate the image from source domain to the target domain. Pix2pix GAN is trained in a supervised manner, in which paired data is required. The drawback of pix2pix GAN is that the paired data is necessary while accessing paired data in real world scenes usually cannot be achieved. Thus, other methods resort to unsupervised learning to train on unpaired data. The CycleGAN [101] builds on the pix2pix GAN. In CycleGAN, it employs cycle-consistency loss on the GAN to ensure the input data and output data contain similar contents. In addition, the CoGAN [51] introduces GAN and variational autoencoders into the Image-to-Image translation framework. By utilizing a weight-sharing strategy, the CoGAN is able to learn a joint representation without any paired image data. A similar work is DiscoGAN proposed by Kim *et al*. [43]. It designs two sets of GANs, where two generators and discriminators learn the mapping between source domain and target domain. Inspired by dual learning from NLP [80], Yi *et al*. [85] proposed DualGAN using conditional GAN [57] to solve cross-domain Image-to-Image trans lation. The DualGAN trains primal and dual GANs to learn domain distribution, which can apply to several image translation applications.

5.4 Proposed Method

In this section, all the details of the proposed UBCDT-GAN will be described. We will first introduce our method, which mainly consists of two networks. The first network Unsupervised Bi-directional Cycle Domain Transfer Network (UBCDTN) aims to

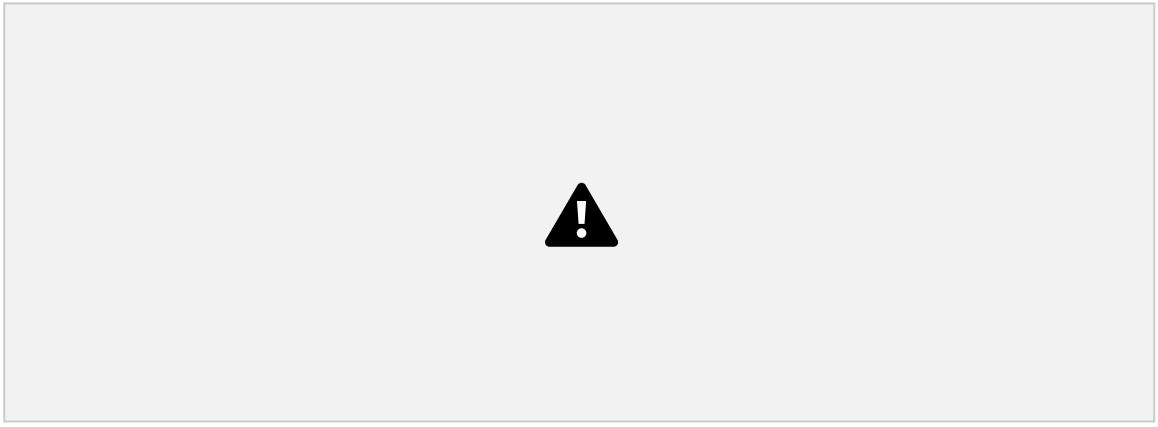

Figure 5.1: The overview of the proposed UBCDT-GAN: In the first stage (left), the green dot rectangle represents Unsupervised Bi-direction Cycle Domain Transfer Network (UBCDTN). The red path indicates the forward cycle module, given the input HR image I^{HR}, I^{LR}

real-like LR image I^{bLR} $_{degraded}$, L^{cyc} $_{degraded}$ is the artificially degraded LR image and $_{idt}$ is produced by G_B. In addition, L^{adv} $_{D_B}$, L^{idt}

age. I^{bLR} $_{G_B}$and L^{percep} $_{F\,E_A}$depicted

$_{real}$ is generated by G_A. I^{bLR}

$_{recon}$ represents the reconstructed im

in the red dotted line represents adversarial loss, identity loss, cycle-consistency loss and cycle-perceptual loss for the forward cycle module. Symmetrically, the blue path shows the backward cycle module, where I^{LR} is given by real-world dataset and syn $_{real}$ thesized LR image I^{LR} $_{syn}$ is generated by G_B. Moreover, the G_A is able to translate $_{syn}$ back to reconstructed real-world LR image L^{LR} $_{recon}$ and generate the identified real-world LR image I^{LR} $_{idt}$. The blue dotted line represents the adversarial loss L^{adv} $_{D_A}$, $_{real}$, cycle consistency loss L^{cyc} identity loss L^{idt}

$_{G_B}$and perceptual loss $L^{percept}$ $_{FE_A}$for backward
cycle module respectively. In the second stage (right), the framework of Semantic En coder guided Super-Resolution Network (SESRN) is depicted in yellow dot rectangle, where it consists of Semantic Encoder *SE*, Generator G_{SR}, Joint Discriminator D_{SR} and Content Extractor φ. There are two paths in the SESRN, where the red path indicates a real tuple and the blue path is a fake tuple. I^{SR} is SR images from G_{SR}. Furthermore, $SE(\cdot)$ denotes the embedded semantics obtained from *SE*. $D(\cdot)$ repre sents the output probability of D_{SR}. $\varphi(I^{HR})$ and $\varphi(I^{SR})$ describe the features learned by φ.

perform domain translation operation on two different domain image datasets. It contains the forward cycle module and backward cycle module, and the pipeline is shown in Figure 5.1. The second network is SESRN consists of Semantic Encoder (SE), Joint Discriminator D_{SR}, Generator G_{SR} and Content Extractor *CE*. The SESRN takes the real-like LR image as input and super-resolve it to the HR image, and the details can be seen in Figure 5.2. Moreover, we will introduce the design of the loss functions. The overview of the proposed as shown in Figure 5.1.