Semantic aware generative adversarial network for image super-resolution

A Thesis

submitted in fulfillment of the requirements for the award of the degree of

DOCTOR OF PHILOSOPHY

in

Electronic and Communication Engineering

By

Shailza Sharma

Regn. No.: 901806016

Under the supervision of

Dr. Vinay Kumar
Associate Professor, TIET

Dr. Abhinav Dhall
Assistant Professor, IIT Ropar

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Department of Electronics and Communication Engineering,
Thapar Institute of Engineering & Technology,
(Deemed to be University),
Patiala (147004), India.

October 2023

Declaration

I, **Shailza Sharma**, hereby certify that the work, which is being presented in the thesis, entitled "**Semantic aware generative adversarial network for image super-resolution**" submitted in Thapar Institute of Engineering and Technology, Patiala in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy in Electronics and Communication Engineering, is an authentic record of my own research work carried out during the period Aug 2018 to June 2023 under the supervision of **Dr. Vinay Kumar** and **Dr. Abhinav Dhall**.

I have also cited the reference about the text(s)/figure(s)/table(s) from where they have been taken. The matter presented in this thesis has not been submitted elsewhere for the award of any other degree or diploma from any institution.

Date: June 13, 2023

Shailza Sharma

Candidate

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Date: June 13, 2023

Dr. Vinay Kumar

Associate Professor

Supervisor

Dr. Abhinav Dhall

Assistant Professor

Supervisor

Abstract

This thesis investigates the field of super-resolution using deep learning methodologies, with a specific focus on Convolutional Neural Networks and Generative Adversarial Networks. The primary objective is to enhance the resolution and quality of low-resolution images by proposing novel architectures and methodologies that address the inherent challenges in this domain.

The first contribution of this thesis is the development of a novel GAN-based architecture for super-resolution. The proposed architecture incorporates a dual-stage upsampling approach for an upscaling factor of 4, utilizing inter and intra residual dense connections. This design enables the model to effectively capture high-frequency texture details in images. Furthermore, the integration of semantic information with the input image enhances the depiction of objects, resulting in visually compelling outcomes. To ensure stable training, spectral normalization is employed in the discriminator architecture

The second contribution of this thesis is the introduction of the Generative Adversarial Based SRINet model. This model obviates the need for linear filters by integrating complex filter structures within the network. Additionally, the architecture incorporates dense skip connections to enhance the network's learning capability while retaining computational efficiency. A progressive upscaling approach is employed to preserve high-frequency components and produce output images with fine texture details.

Furthermore, this thesis presents a novel GAN-based progressive face hallucination network. To generate output images with 3D parametric information, an auxiliary supervision network is utilized, leveraging the shape model of a 3D Morphable Model (3DMM) to generate 2D images with 3D parametric information. Additionally, an autoencoder is proposed to incorporate high-frequency components using high-resolution coefficients of Discrete Cosine Transform (DCT). An Inverse DCT (IDCT) block is introduced within the network to convert frequency domain coefficients to the spatial domain, effectively embedding high-resolution DCT information into the face hallucination network.

Lastly, this work investigates the benefits of incorporating audio signals in the video face hallucination task. Empirical evidence demonstrates that audio signals aid in retrieving lost visual information and maintaining visual consistency across consecutive frames. A novel lip-reading loss, inspired by visual speech recognition, is introduced, enabling the proposed architecture to

generate facial images with fine texture details in areas such as the mouth and lips. Additionally, a frequency-based loss function is incorporated to effectively capture salient frequency features.

Acknowledgment

I am immensely grateful to everyone who has been a part of my thesis journey and contributed to its successful completion. The research work related to this thesis was carried out at the Department of Electronics and Communication Engineering, Thapar Institute of Engineering and Technology, Patiala, India, during the years 2018-2023.

First and foremost, I would like to express my deepest gratitude to my supervisors, Dr. Vinay Kumar and Dr. Abhinav Dhall. I am truly fortunate to have had the opportunity to carry out this research under their guidance. Their invaluable support, encouragement, and life teachings have shaped my journey.

My sincere thanks go to the Head of the Department, Professor Dr. Alpana Agarwal, for funding my work and for her unwavering support. Special appreciation goes to the members of my doctoral committee, Dr. Shailni Batra and Dr. Sanjay Sharma.

I would like to extend my heartfelt thanks to my parents, Parkash Sharma, Santosh Sharma, Tarsem Lal, and Kuldeep Kaur, for their constant love and support. I am grateful to my brothers, Sumit and Vikas, and my friends Swati, Ritika, Akanksha, Sukhmani, Ramneek, Pankaj and Sanjay, your friendship and support have meant the world to me.

Last but certainly not least, I want to express my deepest gratitude to my husband, Vivek Singh Bawa. Your love, understanding, and unwavering support have been my pillar of strength throughout this journey. Thank you all from the bottom of my heart for being a part of this remarkable chapter in my life.

Shailza Sharma

Contents

List of Figures

List of Tables

1 Introduction

1.1 Overview

A digital image comprises small picture elements called pixels. The number of pixels per unit area defines the spatial resolution. In various applications of digital image processing, high spatial resolution images are required. Capturing high-resolution images requires costly imaging sensors. The high density of these sensors indicates a higher spatial resolution. Therefore, to capture images with high-resolution, the image acquisition sensor's density needs to be increased, which in turn increases the hardware cost. In addition, they need higher bandwidth and a large amount of data for transmission. Camera speed and optical blur are other factors that affect the resolution of images. Therefore, low-resolution images could result from hardware limitations, such as a low-resolution camera sensor or a low-bandwidth network connection, or intentionally downscaled to reduce storage or transmission costs. These low-resolution images can suffer from a loss of critical visual details, making them difficult to interpret or analyze. Therefore, the general techniques to enhance the resolution of images and generate high-resolution images are:

1. Increasing sensor density,
2. Increasing the chip size of the capturing device, and
3. Super-resolution.

The first two methods are generally not preferred since they lead to increased cost of the device and higher computational complexity. Therefore, it is preferred to capture an image as a low-resolution image and perform some pre and post-processing on the degraded image to get a high-resolution image. Image super-resolution [1, 2] is the process of upscaling a low-resolution image to a higher-resolution image while preserving or enhancing important details in the image. The goal of image super-resolution (SR) aims to generate a higher-quality image with a higher level of detail and clarity than the original low-resolution image.

SR is an inverse problem [3] where pixels participate in generating unavailable data, and therefore, there is no unique solution for this problem. The SR algorithm depends on the number of input and output images used. In single-image single output (SISO) SR [4], a single low-resolution image is used to produce a high-resolution image. In multiple-input single-output

(MISO) SR [5, 6], multiple low-resolution frames estimate a single high-resolution frame. MISO SR requires fusion stages and high computational complexity during image registration and is not a very popular method. Therefore, single-image SR is preferred over multi-frame super-resolution.

There are several approaches to image super-resolution, including interpolation, reconstruction, and learning-based methods. Interpolation-based methods [7] involve simple algorithms such as the nearest neighbor [8], bilinear, or bicubic interpolation to upsample the low-resolution image to the desired resolution. Reconstruction-based methods [2] involve estimating a high-resolution image from multiple low-resolution images, typically by exploiting the relationship between the images in a Bayesian framework. Learning-based methods [9] involve training a neural network to map low-resolution images to high-resolution ones, typically using deep learning techniques such as Convolutional Neural Networks (CNNs) [10].

Deep learning-based methods have shown promising results in image super-resolution, achieving state-of-the-art performance on benchmark datasets [11]. These methods typically involve training a CNN to learn a mapping function between low-resolution and high-resolution image pairs, using a dataset of paired images to learn the parameters of the network. The trained network can then be used to generate high-quality super-resolved images from new low-resolution images.

Image super-resolution is an important technique for improving the quality and usability of low-resolution images in many applications. With the development of deep learning-based methods, generating high-quality super-resolved images from low-resolution inputs has become increasingly feasible, enabling new applications in fields such as medical imaging, video streaming, and computer vision.

1.1.1 Applications

Image super-resolution has a wide range of applications in various fields, including computer vision, medical imaging, video streaming, and digital photography. The applications that make advantage of image super-resolution methods are listed below.

1. Medical imaging [12, 13, 14]: Super-resolution imaging techniques are used in medical imaging to provide clearer, more accurate images that help doctors and other healthcare workers diagnose and treat a variety of medical disorders.

2. Satellite imaging [15, 16]: Using image super-resolution in satellite imaging provides researchers with more detailed and accurate geospatial data and hyperspectral images, which helps in analyzing and monitoring various environmental and social phenomena.

3. Digital photography [17] : In digital photography, image super-resolution allows photographers to improve the resolution and clarity of digital photos, which is extremely useful in applications like printing or online sharing.

4. Art restoration [18]: Image super resolution technology can assist art historians and restorers in enhancing the quality and resolution of artwork images in order to restore and conserve the cultural legacy for future generations.

5. Video processing [19, 20]: Video satellite imagery is a new technique for earth dynamic observation and has a wide range of uses in environmental fields. Despite its capability of dynamic target detection, it sustains a severe restriction of image quality due to the degradation and compression in its imaging process. In order to improve the quality of the frames in video satellite imagery, super-resolution techniques can be applied.

6. Surveillance and image forensic [21, 22]: Generating a high-quality resolution image has become essential in the forensic field. Video frames of common security surveillance cameras are found to be very low in clarity and degraded with many noises, distortions, blurs, and lousy illumination. So, SR techniques are required to enhance the quality of videos.

7. Face hallucination: Applying super-resolution on faces is known as face hallucination (FH) and is widely required in many image processing applications, such as facial emotion detection [23], pedestrian reidentification [24], facial alignment [25], face recognition [26, 27], face identification [28] and deep fake detection [29].

8. Gaming [30, 31]: By improving video gaming frames' visual quality and resolution, image super-resolution techniques can give players a more engaging and realistic gaming experience.

1.1.2 Challenges

Numerous methods have been proposed to solve SR problem. Still there are various issues that remained unsolved. Following are the challenges that need to be addressed by researchers:

1. Super-resolution algorithm based on reconstruction based methods (edge directed and interpolation based) are fast as comparison to deep learning based algorithms but they oversimplify the SISR problem. They usually yield solutions with over smooth textures and missing high frequency details. Interpolation can lead to blurry images, while reconstruction-based methods can be computationally expensive and require prior knowledge of the image structure. These methods also tend to struggle with handling complex image textures and patterns.

2. Deep learning-based methods face challenges in terms of data availability, model architecture, and computational resources. These methods require large datasets of high-resolution and low-resolution images for training, which can be difficult to obtain in certain domains. Model architecture is also an important factor in achieving high-quality results, and designing optimal architectures can be a challenging task. In addition, deep learning-based methods require significant computational resources for training and inference, which can be expensive and time-consuming.
3. Images with different scales and aspect ratios can pose difficulties in aligning the input and output images, which can lead to distortion and artifacts in the output image.
4. Deep learning models trained on specific datasets can struggle to generalize to new datasets or domains, leading to poor performance in image super resolution.
5. GAN based SR networks have high computational complexity and problem of vanishing gradient arises with increase in depth of network.
6. Most current face super resolution methods rely on two-dimensional facial priors to generate high resolution face images from low resolution face images. These methods are only capable of assimilating global information into the generated image. Still the local features, subtle structural details and depth information is missing in final output image.
7. Quantitative results produced by super resolution methods based on generative adversarial networks have less values as compared to other deep learning based methods.
8. Face super resolution task remains rather challenging in videos in comparison to the images due to inherent temporal consistency issues.
9. Another major issue in present video based SR approaches is the presence of blurriness around the key facial regions such as mouth and lips - where spatial displacement is much higher in comparison to other areas.

1.2 Objectives of dissertation

This thesis work was accomplished with primary focus on the following three objectives:

1. Development of a novel generative adversarial network to improve the quality of state-of-the-art in image super resolution.
2. Incorporate semantic information from the scene into the image super resolution system for better perception of an output image.
3. Quantitative and qualitative analysis of the proposed method.

1.3 Structure of the thesis

Chapter 1 provides an overview of image super-resolution and approaches that are used to perform image super-resolution. Then we discussed its applications and the challenges the research community faces while performing it.

Chapter 2 covers the fundamentals of deep learning, along with a thorough explanation of the various deep neural networks. We also discussed how these deep neural networks help perform image super-resolution.

Chapter 3 provides a detailed literature survey on different approaches to performing image super-resolution on generic images. Next, we covered the most recent developments in face hallucination, an application of image super-resolution. We also covered several methods for carrying out video-based super-resolution.

Chapter 4 presents a novel Generative Adversarial Network based architecture named Residue and Semantic feature-based Dual Subpixel Generative Adversarial Network for image super-resolution. This chapter provides insight into how semantic-based information can help the GAN-based network to generate images with high texture.

Chapter 5 presents our work to generate an efficient GAN-based image super-resolution architecture. Complex filter settings are used in the proposed work to reduce the computational complexity of the model while retaining the quality of images.

Chapter 6 proposes a semantic structural constraint based face hallucination method along with frequency supervision using generative adversarial networks. Extensive experimentation evaluation shows the usefulness of the proposed architecture in the form of state-of-the-art quantitative results.

Chapter 7 presents a video-based face hallucination method using audio-visual cross-modality support. In this work, we also propose a novel loss function to mitigate the blurriness around the mouth region and map the spatial displacement.

Chapter 8 presents the conclusion for the proposed works. We also discuss the future direction of the proposed research in this chapter.

1.4 Contribution of the thesis

The following are the contributions of this dissertation:

1. A novel GAN based architecture is proposed for super-resolution where dual stage upsampling is done for an upscaling factor of 4. In dual upsamping stages, inter and intra residual dense connections are done; making our model capable of sustaining high texture

details of an image. To enhance the objects present in the image, semantic information is merged with the input image; leading to excellent visual results.

2. An efficient GAN based architecture is presented for generic image super-resolution. This model alleviates the use of linear filters and integrates the complex filter structures in the network to approximate most favorable sparse structures. Dense skip connections are introduced in the architecture. This approach increases the learning capability of network while retaining its computational complexity.

3. A novel GAN based progressive face hallucination network is proposed. To generate the final output image with 3D parametric information, proposed model uses a auxiliary supervision network which is compelled to generate 2D images with 3D parametric information using shape model of 3DMM. To incorporate high frequency components in the output image, an auto encoder is proposed which generates high resolution coefficients of DCT.

4. In this work, the semantic relation between audio waves and corresponding visual frames is explored to maintain temporal consistency across the frames of videos. We proposed a novel lip-reading loss inspired by automatic speech recognition to generate video frames with high textural information around the mouth region.

1.5 Publications

1. S. Sharma, and V. Kumar, " An efficient image super resolution model with dense skip connections between complex filter structures in Generative Adversarial Networks, " Expert Systems with Applications, vol. 186, pp. 115780, 2021.

2. S. Sharma, A. Dhall, and V. Kumar, " Frequency aware face hallucination generative adversarial network with semantic structural constraint, " Computer Vision and Image Understanding, vol. 223, pp. 103553, 2022.

3. S. Sharma, A. Dhall, V. Kumar and V. Singh " Dual Stage Semantic Information Based Generative Adversarial Network for Image Super-Resolution, " Indian Conference on Computer Vision, Graphics and Image Processing, 2023 (Accepted).

4. S. Sharma, A. Dhall, V. Kumar and V. Singh " Audio-visual video face hallucination with frequency supervision and cross modality support by speech based lip reading loss, 2023 (Under review).

2 Basics of Deep Learning

2.1 Introduction

Deep learning has a transformative impact on artificial intelligence, leading to remarkable advancements across various applications. Notably, in natural language processing (NLP) [32], computer vision [33], and speech recognition [34], deep learning has brought about unprecedented breakthroughs.

Deep learning refers to a class of machine learning algorithms [35] that are based on artificial deep neural networks (DNN) [36]. Neural networks consist of multiple layers of interconnected nodes, which process input data and make predictions based on learned relationships between the input and output. Unlike traditional machine learning algorithms that rely on hand-crafted features, deep learning algorithms automatically learn high-level abstractions from data, allowing them to achieve remarkable accuracy on complex tasks.

Deep learning exhibits a significant advantage in its capability to acquire hierarchical representations of data, as emphasized in the survey by Dong et al. [37]. In computer vision, for example, deep learning algorithms learn to recognize basic features, such as edges and corners in the lower layers of a neural network, and then build up to recognizing more complex features such as objects and scenes in the higher layers. The capacity of deep learning algorithms to learn hierarchical representations of data has empowered them to surpass traditional machine learning approaches in various computer vision tasks. In tasks like object recognition [38]) and image segmentation [39], deep learning algorithms have exhibited superior performance.

Deep learning possesses another notable advantage in its capacity to effectively handle vast volumes of data, which proves crucial for various applications, particularly in NLP and computer vision. Deep learning algorithms excel in learning from extensive sets of data, including text, speech and images, enabling them to extract meaningful representations that are instrumental in tasks such as language translation and speech recognition and digital image processing.

The achievements of deep learning can be attributed to several critical factors, including the abundance of sizable datasets, the availability of robust hardware resources, and the advancements in sophisticated training algorithms. These factors collectively contribute to the success of deep learning in various applications. The rise of cloud computing and the availability of

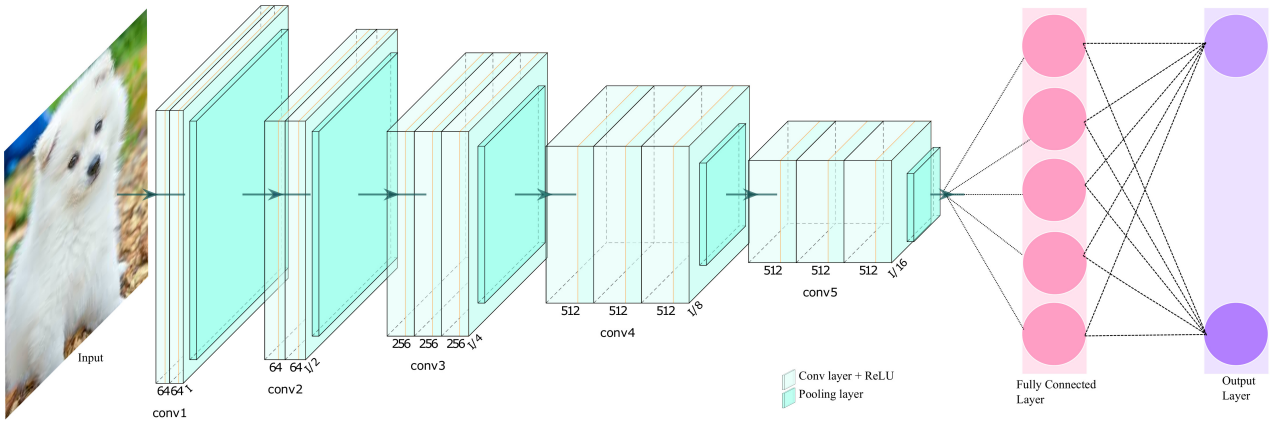


Figure 2.1: Convolutional Neural Network Architecture

GPUs have made it possible to train large deep learning models on massive amounts of data, leading to unprecedented levels of accuracy on a variety of tasks.

The advent of deep learning has a profound and transformative impact on artificial intelligence. It has not only unlocked new possibilities for addressing intricate problems but has also gained extensive adoption across a wide range of applications. With ongoing advancements in hardware capabilities and the continuous development of novel training algorithms, deep learning is poised to maintain a prominent role in shaping the future of artificial intelligence. CNNs [10], generative adversarial networks (GANs) [40], and auto-encoders (AE) [41] are among the DNN architectures that have recently exhibited exceptional performance in various computer vision tasks.

2.1.1 Convolutional Neural Networks

CNNs (refer Figure 2.1) have become a prominent class of DNNs extensively employed in computer vision applications. They have found success in tasks including image classification, object detection, and segmentation [42, 43]. CNNs are designed to process images by extracting meaningful features from them, and they are inspired by the organization and functionality of visual cortex in human brain. CNNs works o the fundamentals of convolution, which involves sliding a filter over an image and computing a dot product between the filter and the local patch of the image. This operation produces a new feature map, which highlights the presence of certain patterns or textures in the image. Convolutional filters are typically small and learnable, allowing the CNN to automatically discover the most relevant features for a given task. CNNs are composed of multiple layers, each of which performs a specific function which are as follows:

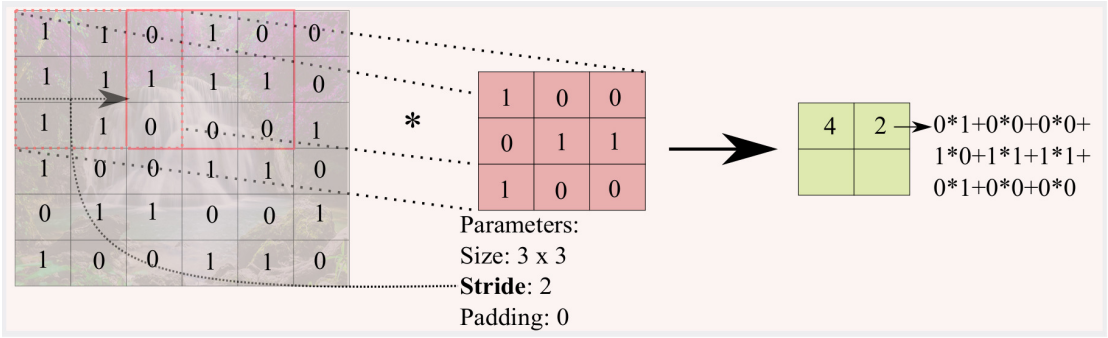


Figure 2.2: Convolution layer operation

Convolution Layer:

Convolutional layers are crucial components within CNNs, serving as the main building blocks. Their primary purpose, as depicted in Figure 2.2, is to apply learnable filters, also known as kernels or feature detectors, to the input data with the objective of extracting high-level features. The convolutional layer achieves this by performing a convolution operation, which entails sliding the filter across the input data and computing the dot product between the filter weights and the corresponding input data at each position. This process yields a set of activation values known as feature maps or output channels, which effectively capture the presence of specific patterns or features within the input data.

During the training phase, the parameters of the filters in a convolutional layer are learned through backpropagation. This involves computing the gradient of the loss function concerning the filter weights and using it to update the weights. This adaptive learning process enables the filters to effectively capture the specific features in the input data and learn increasingly intricate representations as the network delves deeper. Additionally, convolutional layers often incorporate other operations to enhance their performance and mitigate overfitting, thus improving the overall efficiency and effectiveness of the layer, such as;

- Strides: By adjusting the step size of the filter as it moves over the input data, the output feature maps can be made smaller, reducing the number of computations and potentially improving the network's ability to detect small or fine-grained features.
- Padding: To preserve spatial information and prevent edge effects, padding involves introducing additional pixels or values along the edges of the input data. This technique ensures that the output feature maps maintain same size as input data. Using padding, the convolutional layer effectively mitigates any loss of spatial information and prevents undesired artifacts that may arise at the edges of the input data.
- Nonlinear activation functions: Applying a nonlinear function to the output of each filter enables the network to acquire more intricate and expressive representations. This nonlinearity enhances the network's capability to capture complex patterns and relationships

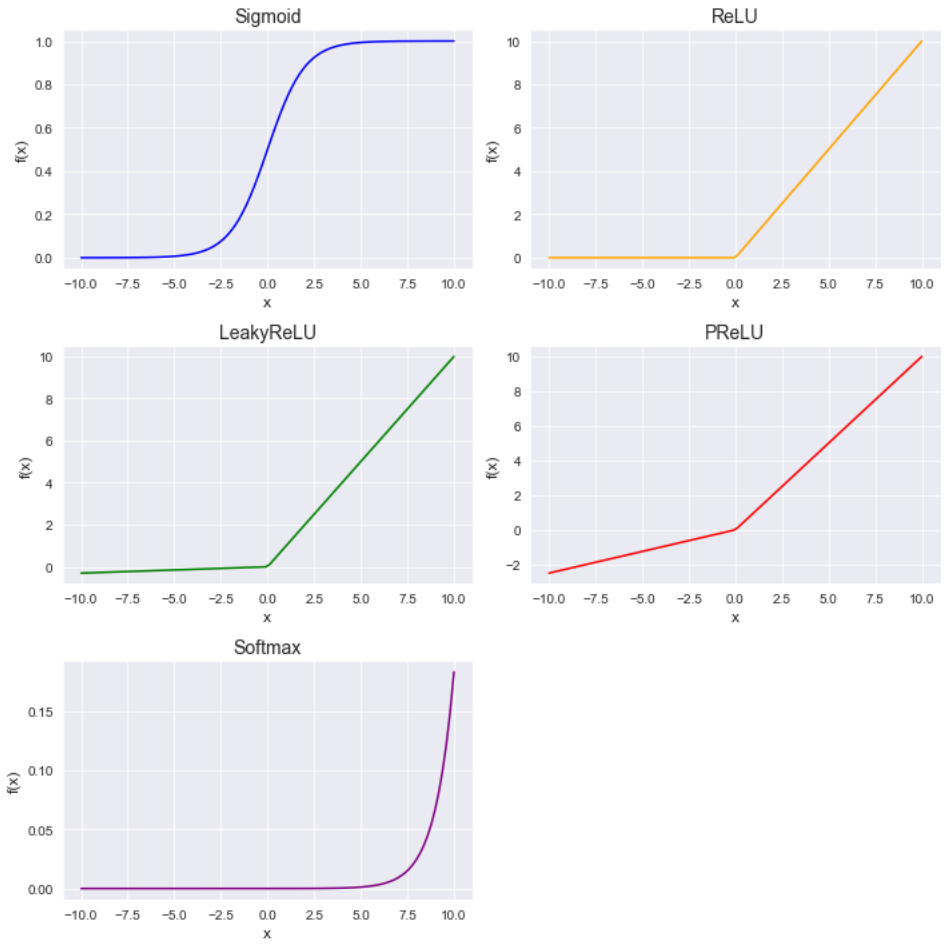


Figure 2.3: Activation Functions

within the data, enabling it to learn and represent more sophisticated features.

Activation Functions:

Activation functions (refer Figure 2.3) are a critical component of CNNs and other types of neural networks. They introduce nonlinearity into the model and help the network learn complex patterns and relationships in the data. Following are some of the commonly used activation functions in convolution layers and their properties.

- Sigmoid Activation Function: This function is defined in eq. 2.1. It has a characteristic S-shaped curve that maps any input value to a value between 0 and 1. It is useful for binary classification problems because it produces a probability output that can be interpreted as the likelihood of the input belonging to the positive class. However, sigmoid activation functions have some drawbacks. First, the gradient of the sigmoid function is very small for large or small input values, which can slow down the learning process during training. Second, the output of the sigmoid function is not centered around zero, which can lead to vanishing gradients and unstable training.

$$f(x) = 1/(1 + exp(-x)) \tag{2.1}$$

- The Rectified Linear Unit (ReLU): ReLU is defined in eq. 2.2. It is one of the most widely used functions in CNNs and other deep learning models because it is simple, computationally efficient, and can mitigate the vanishing gradient problem. The ReLU function is linear for positive input values, which means that it does not saturate or plateau like sigmoid function. This allows the network to learn more quickly and avoid the vanishing gradient problem that can occur with other activation functions. However, the ReLU function is not differentiable at $x = 0$, which leads to unability to calculate the gradient at $x = 0$. Since, all the negative values are always zero, the corresponding weights and biases will never get updated, in turn making the gradient to be zero. This problem is called dying ReLU. There are few variants of ReLU that are designed to increase the further efficacy of existing function and resolve the dying ReLU problem

$$f(x) = max(0, x) \tag{2.2}$$

- LeakyReLU: Dying ReLU problem is resolved by LeakyReLU defined in eq. 2.3. For negative values it adjoins a very small positive slope. Therefore, during back propagation the gradient doesn't become completely zero. α is a very small constant value multiplied with the negative input values.

$$f(x) = max(\alpha x, x) \tag{2.3}$$

- Parametric ReLU (PReLu): It is the advanced version of LeakyReLU, the α value is not fixed initially but learned adaptively during the training according the problem statement.
- Softmax: In CNNs, the Softmax function is commonly employed in the output layer for multi-class classification tasks. This function normalizes an input vector, producing a probability distribution across K classes, where K represents the number of classes. The mathematical formulation of the Softmax function is presented in Equation 2.4, where x_i denotes the ith element of the input vector, and j iterates over all K classes. By applying the Softmax function, the output values represent the probabilities that the input belongs to each of the K classes, facilitating the interpretation of class membership likelihoods.

$$f(xi) = exp(x_i)/sum(exp(x_j)) \tag{2.4}$$

The selection of an activation function in a CNN is influenced by both the problem at hand and the network's architecture. Among the various options, two commonly utilized activation

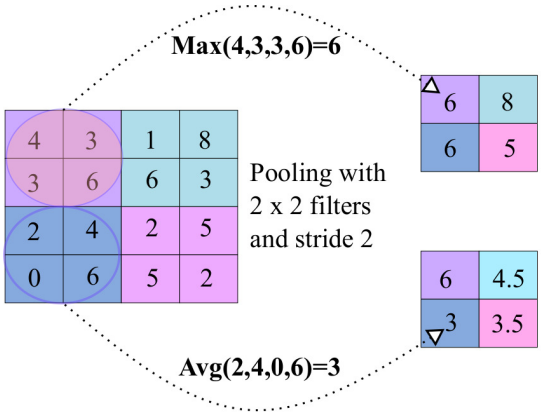


Figure 2.4: Pooling layer

functions are Sigmoid and ReLU. The Sigmoid function is frequently employed for its ability to squash the input into a range of values between 0 and 1, which is advantageous in tasks such as binary classification. On the other hand, the ReLU function, which sets negative values to zero and leaves positive values unchanged, aids in capturing non-linear relationships and is widely used due to its simplicity and computational efficiency. Additionally, the Softmax function is employed explicitly in the output layer of a CNN for multi-class classification problems, enabling the transformation of the network's output into probabilities representing class membership likelihoods.

Pooling Layer

A pooling layer (refer Figure 2.4) is used to reduce the spatial dimensions (i.e., height and width) of the input volume while retaining the most important features. They are commonly utilized following convolutional layers to progressively downsize the feature maps, thereby aiding in the prevention of overfitting. The pooling operation typically involves a sliding window moving across the input volume and computing some function on the subregions of the input. The two most prevalent types of pooling are max pooling and average pooling.

- Max pooling: This pooling operation computes maximum value within each subregion of the input. It effectively reduces the dimensionality of the input volume by retaining only the largest value in each subregion, and is useful for detecting important features and maintaining spatial invariance.
- Average pooling: It computes the average value within each subregion of the input. It is a simpler operation that can help to reduce overfitting by reducing the dimensionality of the input volume.

The size of the pooling window, also called the pool size, is a hyperparameter that can be adjusted to control the degree of pooling. A larger pool size will result in more aggressive

pooling and greater spatial reduction, while a smaller pool size will lead to less reduction and more detail retention.

Fully connected (FC) layer

In the final classification stages, the FC layer assumes a vital role by converting the high-level representations acquired from earlier layers into suitable output class probabilities or predictions. While convolutional and pooling layers are primarily dedicated to extracting local features and reducing spatial dimensions, the FC layer focuses explicitly on mapping the learned features to their corresponding output classes.

A FC layer connects every neuron in the previous layer to every neuron in the current layer, forming a dense matrix of weights that are learned during training. The input to a FC layer is typically flattened output of the preceding convolutional or pooling layers, which has been reshaped into a vector.

In an FC layer, each neuron performs a weighted sum of the input activations and a bias term. This computation is followed by applying an activation function, to introduce non-linearity to the neuron's output. The resulting output from each neuron represents a specific feature or class probability. Ultimately, the FC layer produces a probability distribution across the potential output classes.

Adjusting the number of neurons in the FC layer is a hyperparameter that influences the network's complexity and the number of possible output classes. Typically, the FC layer contains a larger number of neurons compared to the preceding convolutional or pooling layers. This arrangement allows the layer to capture more high-level features and abstract representations, enabling the network to make informed predictions.

To ensure the output probabilities are normalized and sum up to one, a softmax activation function is commonly applied after the FC layer. This normalization step facilitates interpreting the output probabilities as a probability distribution. These probabilities are then utilized for predicting the class of the input image. The FC layer plays a critical role in a CNN, as it enables the network to learn and utilize high-level representations of input images for accurate classification.

2.1.2 Generative Adversarial Networks (GANs)

GANs, as depicted in Figure 2.5, belong to a class of deep learning models that comprise two interconnected neural networks: a generator network and a discriminator network. These models, introduced by Ian Goodfellow in 2014 [40], have gained significant popularity in generative modeling.

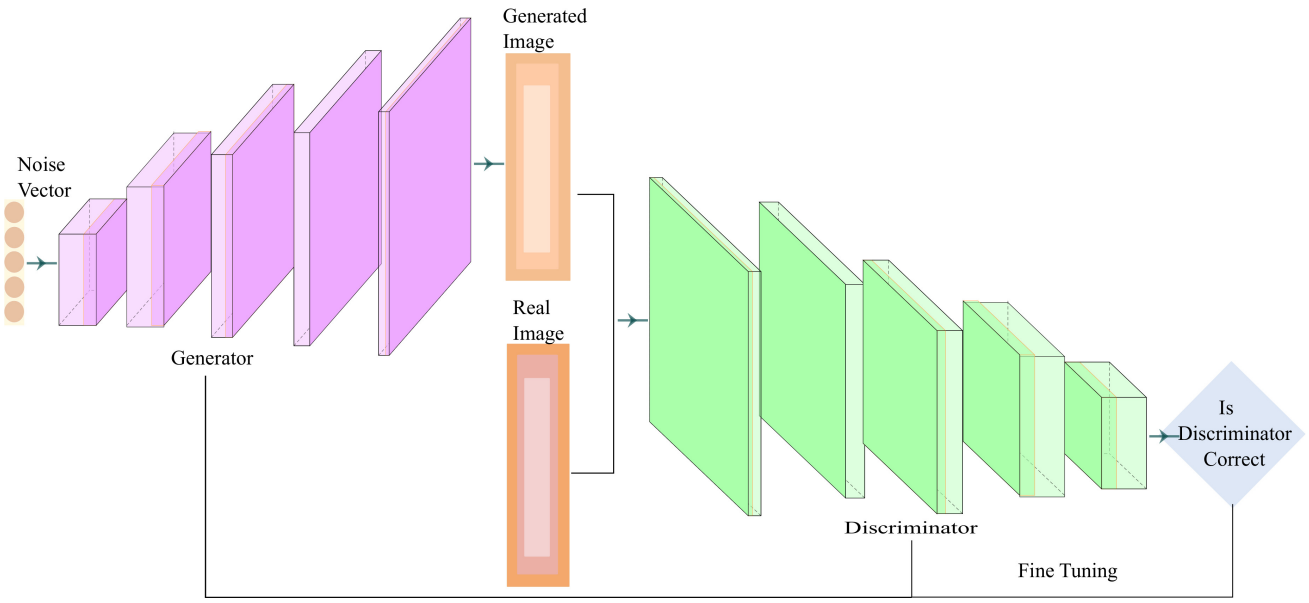


Figure 2.5: Generative Adversarial Network Architecture

The generator network plays a crucial role in GANs. It inputs a random noise vector and synthesizes artificial samples that closely resemble actual data. On the other hand, the discriminator network is responsible for distinguishing between actual samples from the training set and the synthetic samples generated by the generator. It performs binary classification, classifying input samples as either natural or synthetic.

By training these two networks together in an adversarial manner, GANs aim to achieve a competitive dynamic. The generator strives to produce increasingly realistic samples that can fool the discriminator, while the discriminator aims to accurately discriminate between actual and synthetic samples. Through this iterative process, GANs learn to generate high-quality and visually convincing synthetic data that closely resemble the actual data distribution.

In GANs, the loss function comprises two components: generator loss and discriminator loss. These terms are combined to guide the training process. The generator loss, as shown in equation 2.5, evaluates the generator's effectiveness in deceiving the discriminator by producing synthetic samples classified as real. A popular choice for the generator loss is the binary cross-entropy (BCE) loss. This loss is computed by comparing the discriminator's predictions for the generator's samples with a vector of ones, indicating that the samples should be classified as real. The BCE loss quantifies the discrepancy between the discriminator's outputs and the desired classification, encouraging the generator to generate more convincing samples that resemble the real data distribution.

$$-[log(1 - D(G(z)))] \tag{2.5}$$

The variable z represents a random noise vector that acts as input to the generator. The dis-

criminator, on the other hand, evaluates the generated sample produced by the generator and produces an output denoted as $D(G(z))$. The negative sign is used to convert the minimization problem of the generator loss into a maximization problem, which aligns with the minimization problem of the discriminator loss.

The discriminator loss quantifies the discriminator's ability to accurately differentiate between authentic and fake samples. A frequently employed approach for the discriminator loss is the BCE loss, as depicted in equation 2.6. It entails calculating the cross-entropy between the discriminator's predictions for both real and synthetic samples and their respective target values, which are represented by vectors of ones and zeros. This loss function guides the training of the discriminator, encouraging it to make accurate distinctions between real and synthetic data points.

$$-[log(D(x)) + log(1 - D(G(z)))] \tag{2.6}$$

In the equation provided, the variable x represents a real sample drawn from the underlying data distribution, while $D(x)$ represents the discriminator's prediction for that particular sample. The first term in the equation denotes the loss incurred by the discriminator when it incorrectly classifies a real sample as fake, while the second term signifies the loss incurred when it erroneously classifies a synthetic sample as real.

The ultimate objective of training a GAN is to reach a state of Nash equilibrium, a concept originating from game theory [44]. In this equilibrium, the generator is capable of producing synthetic data that is virtually indistinguishable from real data, and the discriminator becomes incapable of accurately discerning between real and synthetic samples. This equilibrium state is desirable as it signifies that the generator has successfully captured the essential characteristics of the real data distribution, leading to the generation of highly realistic synthetic data.

Throughout the training process, the generator network improves its ability to generate realistic samples that can deceive the discriminator, while the discriminator network becomes more adept at discerning between genuine and fake samples. Consequently, the final generator network can generate new data that shares similarities with the training set, although it may not be an exact replica.

Applications

GANs have found extensive applications across diverse domains, including:

- Image and video synthesis [45]: GANs have demonstrated the ability to generate highly realistic images and videos, making them valuable in computer vision, gaming, and entertainment.

- Style transfer [46]: GANs can be employed to transfer the visual style of one image to another, enabling the creation of unique and artistic images that combine the content of one image with the style of another.
- Data augmentation [47]: GANs are utilized for generating synthetic data samples that can supplement the training set, improving machine learning models' performance by providing additional diverse and representative data points.
- Anomaly detection [48]: GANs can be used to identify anomalies or outliers in data, which can be useful in fraud detection, security, and medical diagnosis.
- Drug discovery [49]: GANs have been used to generate new and diverse molecular structures, aiding in the discovery of new drugs and materials.
- Text-to-image synthesis [50]: GANs can be used to generate realistic images from text descriptions, which has applications in gaming, virtual reality, and product visualization.
- Super-resolution imaging [51]: By leveraging GANs, it become possible to enhance the resolution of low-quality images, thereby improving the overall quality of various types of images such as medical imaging and satellite imagery.
- Generative music [52]: GANs can be used to generate new and novel pieces of music, with applications in the music industry and entertainment.

Challenges

Although GANs have shown impressive results in generating realistic data samples, they still face several challenges. Here are some of the main challenges of GANs:

- Mode collapse: This occurs when the generator learns to produce a limited set of samples that fool the discriminator, rather than generating diverse and high-quality samples from the full data distribution.
- Training instability: The training process of GANs can be unstable and difficult to converge, resulting in generator and discriminator models that oscillate and fail to produce high-quality output.
- Evaluation metrics: It is challenging to evaluate and compare the performance of GANs, as traditional metrics such as accuracy and loss may not accurately capture the quality of the generated samples.
- Hyperparameter tuning: GANs' performance is significantly influenced by hyperparameters like learning rate, batch size, and network depth. However, effectively tuning these hyperparameters can be a challenging task.

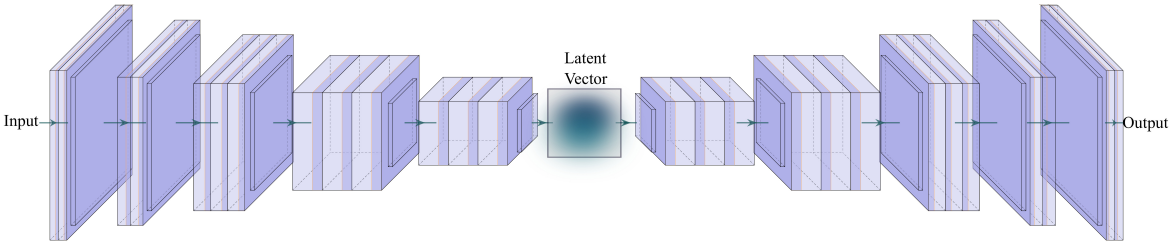


Figure 2.6: Auto-encoder architecture

- Data complexity: GANs face difficulties in generating high-quality samples from complex data distributions, such as those encountered in natural language processing and medical imaging tasks.
- Data efficiency: GANs require a large amount of high-quality training data to learn the underlying data distribution and produce high-quality output, which may not always be available.
- Generalization: GANs can exhibit overfitting behavior, where they perform well on the training data but struggle to generalize effectively to new, unseen data samples.

2.1.3 Auto-encoders (AEs)

AEs, depicted in Figure 2.6, are DNNs used for unsupervised learning to discover data representations [53]. They comprise an encoder network responsible for compressing the input data into a lower-dimensional representation and a decoder network responsible for reconstructing the original input data using this compressed representation

The objective of an AE is to acquire a concise and effective representation of the input data, serving purposes like data compression, denoising, and data generation. To achieve this, the AE is trained to minimize the dissimilarity between input data and reconstructed output data, thereby learning a robust representation.

Autoencoders are typically trained using backpropagation, where the gradient of the reconstruction error is backpropagated through the network to update the weights. There are several types of autoencoders, including the standard autoencoder, denoising autoencoder, and variational autoencoder.

- Standard Autoencoder [54]: The standard AE, as the most basic type of autoencoder [54], comprises an encoder and a decoder network. In this configuration, the encoder network transforms the input data into a lower-dimensional representation, while the decoder network reconstructs the original input data from the representation.
- Denoising Autoencoder [55]: The denoising autoencoder is a type of autoencoder that is designed to remove noise from the input data. During training, the denoising autoencoder

is presented with noisy input data and is trained to reconstruct the original clean data. This approach can be used for tasks such as image denoising, speech denoising, and text denoising.

- Variational Autoencoder [56]: The variational AE is a specific type of autoencoder that adopts a probabilistic approach to capture a concise and effective representation of the input data. Unlike traditional AEs, the variational AE's encoder network maps the input data to a distribution within the latent space, and the decoder network reconstructs the original input data from the latent space. The training objective of the variational AE involves optimizing a lower bound on the log-likelihood of the input data, promoting the model to learn a concise representation.

Applications

AEs find utility across multiple domains, with a variety of applications as outlined below:

- Image and video compression [57]: AEs are used to compress large images and videos while retaining their essential features, leading to efficient storage and transmission.
- Denoising [58]: AEs are utilized to remove noise from images and videos, leading to improved quality and better visualization.
- Anomaly detection [59]: AEs identify anomalies or outliers in data, which can be useful in fraud detection, security, and medical diagnosis.
- Dimensionality reduction [60]: AEs reduce the dimensionality of high-dimensional data, making it easier to visualize and analyze.
- Feature extraction [61]: AEs can be used to extract important features from data, which can be useful in machine learning and pattern recognition.
- Generative modeling [62]: AEs generate new and novel data samples, with applications in art and design.
- Recommendation systems [63]: AEs can be used to model user preferences and recommend products or services based on user behavior.
- Natural language processing [64]: AEs encode and decode text data, with applications in machine translation, summarization, and question-answering systems.

Challenges

Some of the main challenges of AEs are as follows:

- Overfitting: AEs suffer from overfitting, leading to poor generalization performance on new data samples.

- Lack of interpretability: The encoded features learned by AEs may not be easily interpretable by humans, which can limit their usefulness in some applications.
- Data efficiency: AEs rely on a substantial volume of high-quality training data to effectively learn the underlying data distribution and generate high-quality output. However, obtaining such data may only sometimes be feasible or readily available.
- Hyperparameter tuning: AEs depend highly on the appropriate selection of hyperparameters, including batch size, optimizer, learning rate, and network depth. Tuning these hyperparameters effectively can be a challenging task.
- Computational efficiency: AEs can be computationally expensive to train, especially for large and complex data sets.
- Reconstruction quality: The fidelity of the reconstructed output is influenced by the quality of the encoded features, which may be constrained by factors such as the model's capacity or the complexity of the data distribution
- Handling missing data: AEs struggle to handle missing or incomplete data, which can be common in real-world applications.

2.2 Importance of deep learning in SR

Deep learning has transformed the field of image SR by effectively generating HR images from LR inputs. Unlike traditional methods that relied on manual feature engineering and interpolation techniques, deep learning, specifically deep CNNs, enables the extraction of hierarchical representations to capture intricate LR-HR relationships. Deep learning models excel in processing large datasets and have shown remarkable performance in diverse image SR tasks, leveraging abundant paired LR and HR image data and the computational capabilities of GPUs.

2.2.1 CNNs and GANs for image super resolution:

CNNs have exhibited remarkable proficiency in capturing intricate spatial patterns within images, leading to significant advancements in image SR [65]. However, the utilization of deep CNN architectures can be hindered by the issue of vanishing gradients, which impedes convergence and limits their ability to capture complex spatial structures and preserve fine-grained details. In order to address this limitation, the concept of residual learning [66] has been introduced, enabling more efficient convergence and mitigating the challenges associated with vanishing gradients in deeper CNN architectures. Residual learning is a mathematical framework that enhances the training of CNNs by introducing residual mappings. Denoting the input

to a specific layer as x and its corresponding output as $F(x)$, traditional learning aims to directly map x to $F(x)$. However, in residual learning, the focus shifts towards learning the residual mapping $\mathcal{R}(x)$, which represents the discrepancy between the desired output, denoted as $H(x)$, and the current output $F(x)$. Mathematically, the residual mapping is defined as $\mathcal{R}(x) = H(x) - F(x)$. To derive the final output, the residual mapping is incorporated by adding it back to the input x, yielding $H(x) = F(x) + \mathcal{R}(x)$. This formulation allows the network to primarily concentrate on learning the residual components rather than constructing the entire mapping from scratch. By leveraging skip connections that enable the direct flow of information, the network becomes proficient in capturing and assimilating the residual information.

Despite the effectiveness of CNNs [67, 68], their inherent difficulties in accurately representing complex spatial structures and preserving high-frequency details have been surpassed by the superior performance of GANs in generating high-quality super-resolved images [69]. GANs operate through the interplay of a generator network and a discriminator network, engaging in an adversarial training process that produces outputs closely resembling authentic HR images, thereby achieving heightened levels of realism. GANs excel in faithfully reproducing intricate high-frequency details, including intricate textures and sharp edges, which pose challenges for CNNs in faithfully capturing and representing them.

A pivotal advantage of GANs lies in their utilization of perceptual loss, which capitalizes on the comprehensive features extracted from pre-trained CNNs, such as the widely employed VGG network. By considering perceptual quality beyond mere pixel-level disparities, GANs effectively encompass the overall structure and content of the image, resulting in visually compelling and realistic super-resolved images.

Moreover, GANs possess an inherent generative capability, allowing them to acquire complex mapping from LR images to HR images without necessitating explicit, task-specific training. This intrinsic adaptability and flexibility render GANs an appealing choice for various image SR scenarios, surpassing the limitations associated with CNN-based methodologies.

3 Literature Review

3.1 Generic image super-resolution

Deep learning has become a popular approach for solving the SR problem. The literature survey focuses on using CNNs and GANs for image super-resolution.

3.1.1 Different Upscaling techniques for SR

There are different type of upscaling technique to achieve the desired resolution. Most popular approaches are as follows:

1. Bicubic Interpolation: Bicubic interpolation creates a new pixel value by calculating a weighted average of the pixels around it. It is frequently employed for upscaling an image's resolution by a factor of 2, 3, or more. Unlike linear interpolation, which primarily considers the four closest neighboring pixels, it employs a more intricate interpolation method. Since the closest 16 pixels are taken into account, the bicubic interpolation formula is more precise than linear interpolation. While, bicubic interpolation is a well established technique for image upscaling, it has some limitations compared to deep learning based upscaling methods. Since it is a fixed interpolation method that uses a mathematical formula to calculate new pixel values based on surrounding pixels, it cannot learn complex patterns in the data, limiting its ability to produce high-quality upscaled images.
2. Deconvolution Layer: The deconvolution layer [70] also referred to as transposed convolution, serves as the inverse process of convolution and is employed to enhance the resolution of feature maps. It encompasses two primary stages: upsampling and convolution. During the upsampling phase, the spatial dimensions of the feature maps undergo expansion via the insertion of zeros or the utilization of learnable upsampling methodologies. Subsequently, convolution is performed on the upsampled feature maps using learned filter kernels, generating intermediate feature maps. Ultimately, these intermediate feature maps are combined and processed to reconstruct the final output, enabling the capture of more intricate details by incorporating information from neighboring pixels. However, deconvolution layers can also introduce artifacts, such as checkerboard patterns, if they

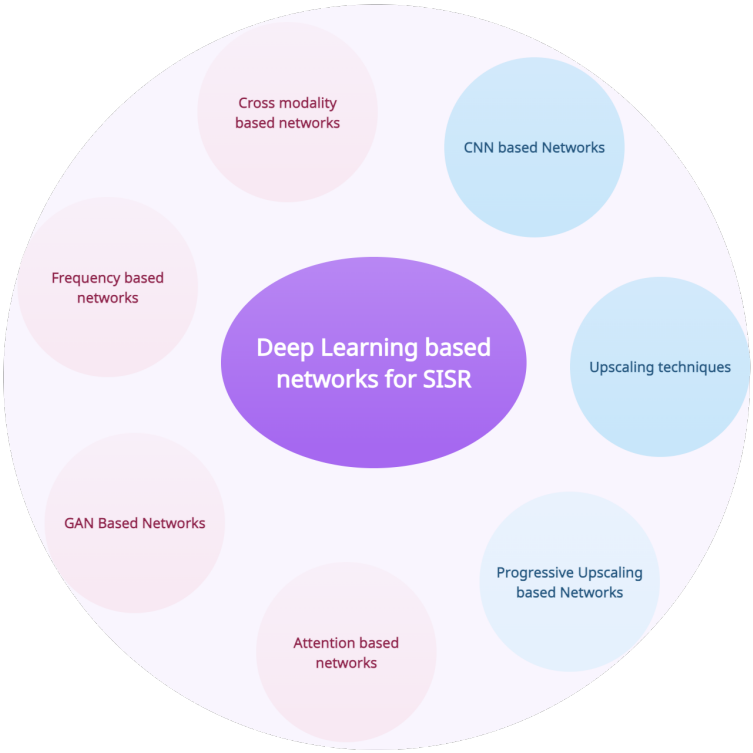


Figure 3.1: Different approaches for image super-resolution

are not used correctly. This is because the insertion of zeros during upsampling can lead to overlaps in the convolutions, creating these artifacts.

3. Sub-pixel layer : In order to mitigate the artifacts problem sub-pixel layer [68] is used. The sub-pixel layer is descendant of the convolution layer.

To derive a sub-pixel layer, U (known as the upscaling factor) is added to the convolution layer. It periodically shuffles the elements of tensor to rearrange elements with shape of super-resolution image. When input image with dimensions $W \times H$ (width and height) is passed through the convolutional layer; it produces the feature map with feature count defined below:

$$F_{count} = \widetilde{N} \times W \times H \tag{3.1}$$

here, $\widetilde{N}$ represents number of feature maps in convolution layer. Eq 3.2 represents the feature maps in sub-pixel layer.

$$\widetilde{N} = (N \times U^2) \tag{3.2}$$

Hence, the final feature count is represented in Eq3.1

$$F_{count} = \widetilde{N} \times U^2 \times W \times H \tag{3.3}$$

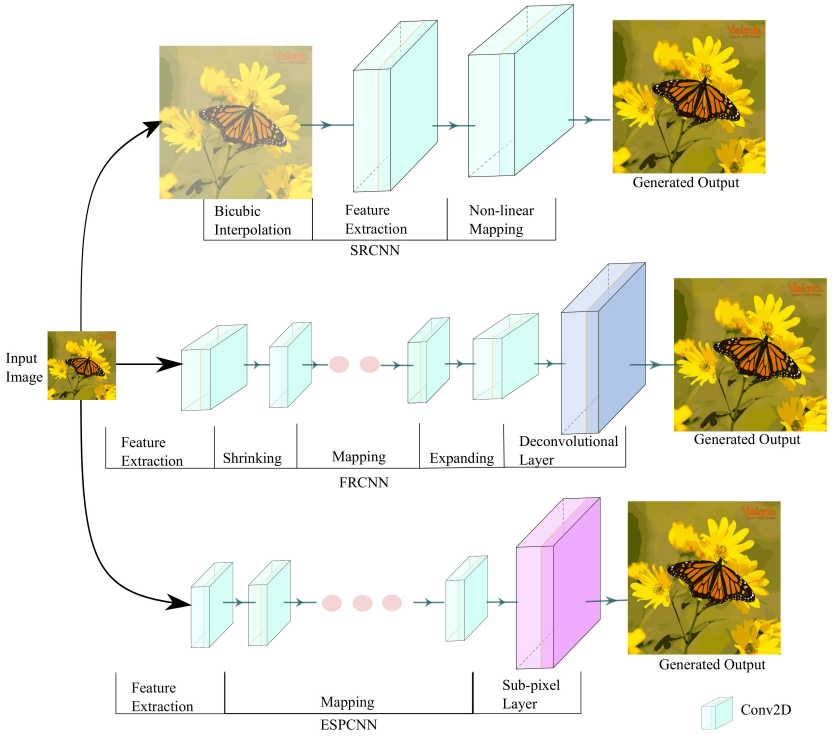


Figure 3.2: Shallow CNN architectures for SR

The sub-pixel layer rearranges the elements of the output tensor to produce shape $N \times (W \times U) \times (H \times U)$ from original shape $((N \times U^2) \times W \times H)$. Hence, unlike deconvolution layer , sub-pixel layer does not induce any new elements in the matrix. It lead to much better performance in SR task. The sub-pixel layer possesses either one or three channels corresponding to grayscale or colored images, respectively. The upscaling factor (U) is pivotal in the model, establishing a connection between the LR and HR feature spaces. It serves as a crucial parameter that influences the mapping of LR to HR features, thereby impacting the overall performance and quality of the model.

3.1.2 CNN based SISR

3.1.2.1 Shallow CNN architectures

The pioneering pre-upsampling SR framework utilizing CNNs, named SRCNN, was initially introduced by Dong et al. [67] in their seminal work. In this framework, the CNN architecture is fed with the bicubic-interpolated variant of the LR image, which serves as input. Through an end-to-end learning process, the CNN learns to establish a mapping between the given input and the subsequent generation of a HR image. This framework showcases the efficacy of employing CNNs for SR tasks, particularly in the context of pre-upsampling.

To speed up the training mechanism in SRCNN architecture, Dong et al. [71] presented the

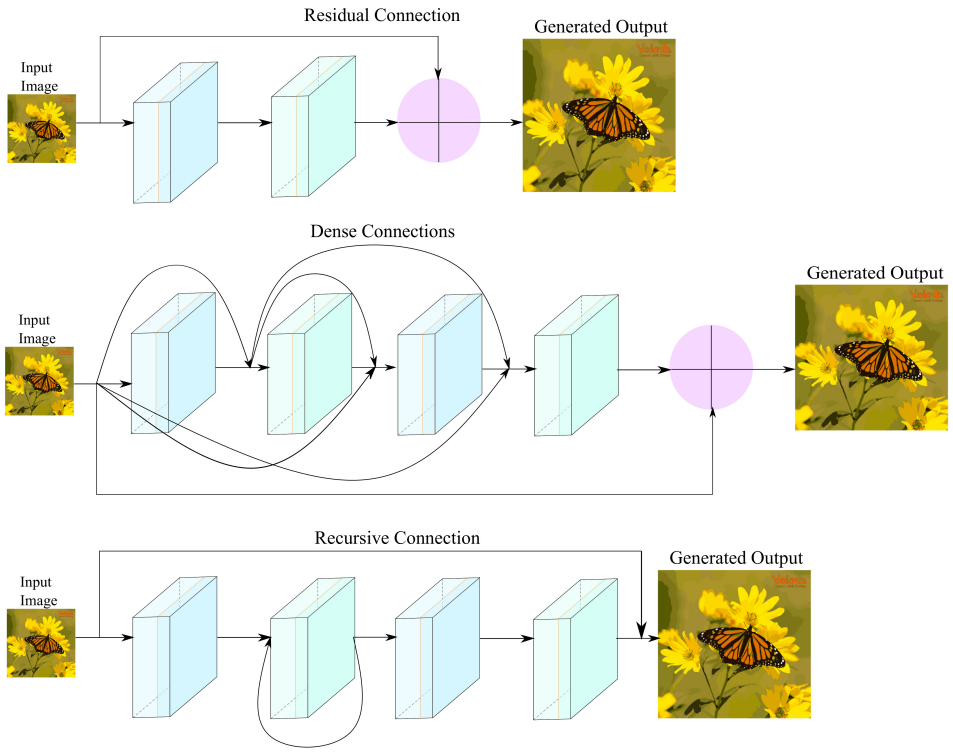


Figure 3.3: Different types of connections present in deep CNN architectures for image SR.

post-upsampling SR framework known as FSRCNN, with shrinking (to reduce the feature dimensionality before mapping) and expanding layer. For upsampling, transposed convolution layer is used.

In work conducted by Shi et al. [68], a CNN architecture named efficient sub-pixel CNN (ESPCN) was introduced. The main objective of this architecture was to facilitate real-time SR by representing extracted features in a LR space. To achieve this, the authors proposed a subpixel layer incorporating an array of upscaling filters for each feature map. This subpixel layer played a crucial role in the upscaling, transforming the image from LR to HR space.

All the aforementioned approaches rely on shallow CNN architectures, which exhibit limited feature learning capabilities. Consequently, these architectures yield suboptimal results in SR tasks. To address the limitations associated with shallow architectures, researchers have turned their attention to exploring deep CNN architectures for SR problems.

3.1.2.2 Deep CNN architectures

With the help of residual learning and optimum learning rate parameters Kim et al. [66] presented a CNN architecture with depth 20 for image SR. VGG-net is the motivation behind this architecture and it resulted in significant increase in the performance of the architecture as compared to the shallow architectures.

The vanishing gradient problem emerges when deep CNNs contain more layers. In order to mitigate this issue, Kim et al. [72] proposed a novel approach in deep CNN architectures, incorporating skip and recursive connections. These connections serve as mechanisms to address the vanishing gradient problem by allowing information to bypass certain layers and enable direct feedback loops.

To overcome the long-term dependency problem in deep CNN, an adaptive learning architecture is proposed by Tai et al. [73] for deep networks. In this architecture, Gate unit stores the output of the previous state and decides how much of this stored value will be used to produce an output by adaptive learning procedure.

To further improve the performance of deep CNN based architectures, a ResNet based architecture is proposed with a series of residual blocks with increasing numbers of filters [74]. This network also incorporates a number of advanced techniques, including residual scaling and a high-pass filter to enhance image details.

Ledig et al. [69] introduced a deep residual network (ResNet) architecture for SR known as SRResNet. This network is a modification of the ResNet architecture that incorporates skip connections to allow information to flow directly between layers, thus reducing the risk of vanishing gradients. The proposed architecture comprises numerous residual blocks, each composed of two convolutional layers with a size of 3×3. It incorporates batch normalization, which normalizes the activations within each mini-batch and ReLU activation function. In addition to these residual blocks, SRResNet incorporates an upsampling block that employs a 3×3 convolutional layer to increase the spatial resolution of the input feature maps twofold. Subsequently, a sub-pixel layer is utilized to reorganize the channels of the convolutional layer, generating the intended HR image.

Since the architectures mentioned earlier only generate images for a single upscaling factor, Li et al. [75] presented a deep CNN network with a multi-scale approach. This network takes a LR inputs images and generates intermediate HR images, each with a different scale factor. The final HR output is then generated by combining the intermediate images in a weighted manner.

In order to gain insights into the degradation process of images from HR to LR and investigate the interdependence between the input and target images, Haris et al. [76] proposed a DNN architecture designed for SISR. The architecture comprises four distinct modules: feature extraction, back-projection, upsampling, and reconstruction. Of particular significance is the back-projection module, which serves as the central component of the architecture and engages in an iterative procedure to refine the feature maps. Refined feature maps through the upsampling and reconstruction modules are utilized to obtain the SR image.

[77] presented a SR method consisting of multiple recursive sub-networks, each of which uses a feature extraction module followed by a feature reconstruction module. The feature reconstruction module includes a residual dense block and a deconvolution layer for upsampling. This

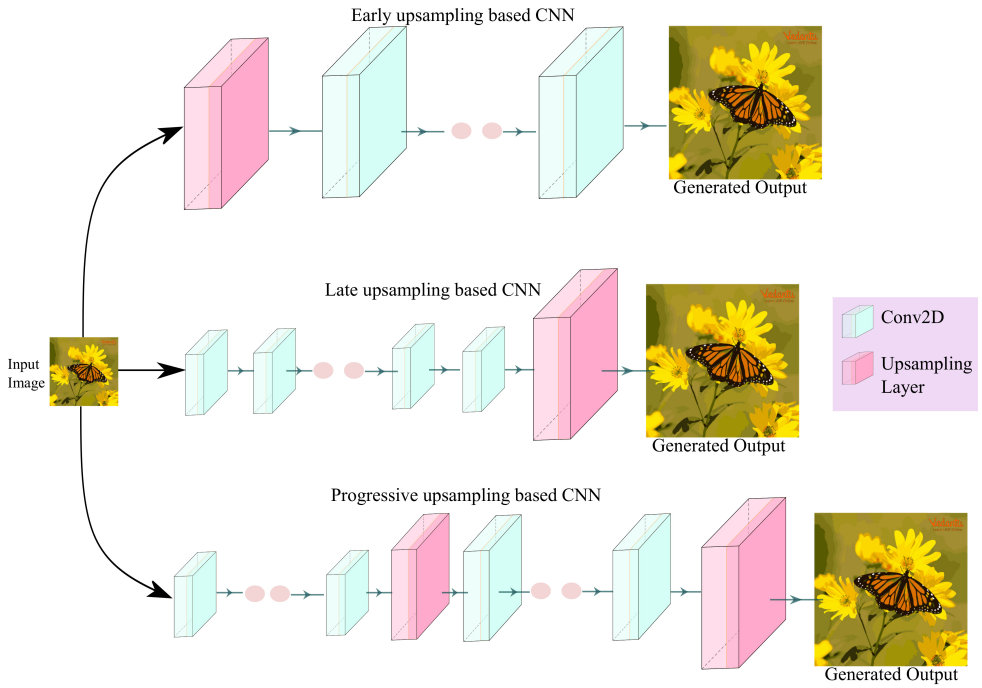


Figure 3.4: Upscaling based network architectures for CNN

network also uses a feedback mechanism to iteratively refine the output image.

To achieve good generalized performance on images with varying scales and contents, fusion model of residual dense connections and error feedback system is utilized by [78]. The proposed model encompasses several stages, where each stage incorporates a feedback connection to iteratively refine its prediction by leveraging the knowledge acquired in the preceding stage. This iterative process facilitates progressive improvement and enhances the accuracy of predictions as the model iteratively learns and integrates information from previous stages.

3.1.3 Progessive upscaling based networks

In their work, Lai et al. [79] introduced LapSRN. This SR architecture leverages the concept of a Laplacian pyramid to reconstruct the HR output from an LR input progressively. The Laplacian pyramid comprises a set of downsampled images alongside their corresponding residual images, representing the difference between each image and its corresponding HR counterpart. These residual images are generated through an LR to HR mapping network, which takes the downsampled image as input and produces a residual image. This residual image is then added to the upsampled image from the previous level of the pyramid. By adopting the Laplacian pyramid framework, the network achieves a gradual reconstruction of the HR output from the LR input.

The task of generating high-quality outputs for significant upsampling factors continues to pose a challenge. In order to address this issue, a pioneering approach known as ProSR [80] has

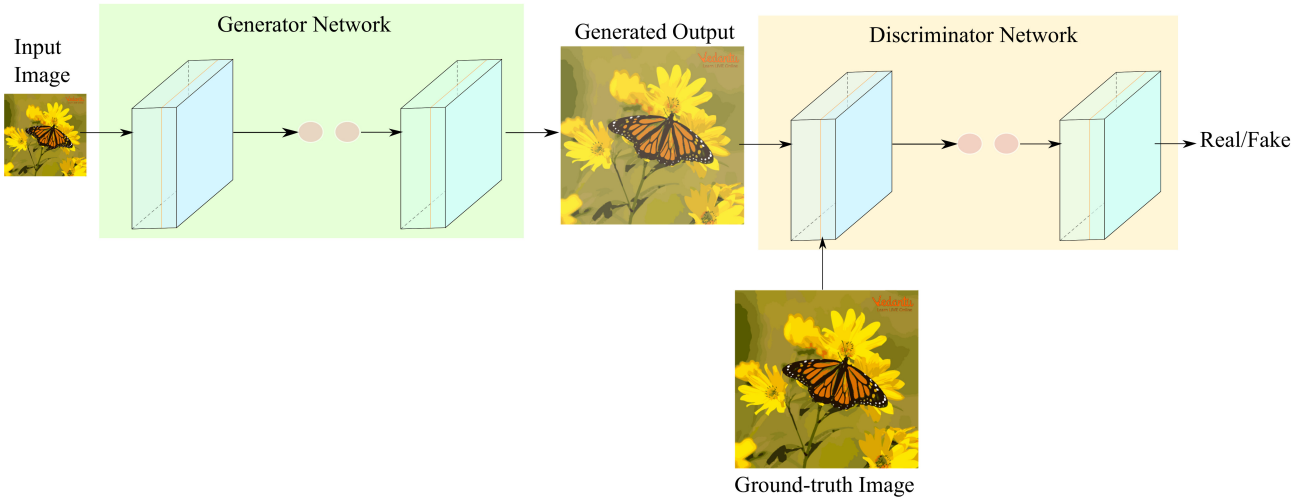


Figure 3.5: Generalized GAN based architecture for SR

been introduced, incorporating a progressive architecture and training strategy. The network systematically increases the resolution of images in incremental stages, leveraging a curriculum learning methodology to optimize learning efficiency. Moreover, an innovative GAN called ProGanSR has been proposed, prioritizing the enhancement of photorealism through the utilization of a progressive multi-scale descriminator.

Chudasama et al. [81] introduced ComPrESRNet, a progressive and compact network designed for superior SISR. The proposed methodology encompasses a series of stages, each employing a compact residual block based architecture to generate HR images with increasing levels of intricacy. The network integrates residual-in-residual blocks and dense connectivity to enhance feature learning while mitigating the issue of vanishing gradients. Furthermore, a feedback mechanism is introduced to augment the network's capability to refine its predictions based on previously generated images.

3.1.4 GAN based SR

Loss functions, like MSE, rely on pixel-wise difference. This pixelwise dependency diminish their capability to retrieve high frequency features, generating images with very low visual quality [82, 83]. To solve this problem, GAN based architecture is proposed for SR framework by Ledig et al. [69], where they integrated two loss functions, namely content and adversarial loss. Rather than relying on pixel-wise differences, the authors computed feature-wise differences to generate SR images.

SRGAN architecture and loss functions are reanalyzed by [84]. They used a generator architecture based on the Residual-in-Residual Dense Block (RRDB) design, incorporating residual connections within each RDB block to learn more efficient image representations. Additionally, the generator employed a Feature Fusion Module (FFM) to combine the outputs of the RRDB

blocks, resulting in the generation of the final HR image. Furthermore, discriminator architecture included a feature matching loss, which quantified the disparity between the discriminator's feature representations of the generated and authentic HR images, thereby improving the quality and realism of output images.

Zhu et al. [85] proposed a discriminator architecture for GAN-based SISR framework utilizing the Wasserstein distance. In their work, the generator architecture was designed with dense skip blocks, which aimed to enhance learning efficiency by facilitating information flow across different layers. Additionally, transition blocks were introduced, employing 1×1 convolution layers to deepen the network while reducing the overall number of parameters.

[86] argued that the LR images used in training a GAN are inadequate to represent the real world LR images. So, they introduced a combined network where at first, image learns a real world degradation process and then the degraded image is converted to HR image.

To resolve issue of misalignment for LR images with small size, discriminative generative architecture is proposed by Yu et al. [87]. To accurately mimic the features of face, a spatial support is provided across with the input image as a high-frequency residual and $l2$ regularization term is added to loss function. The added loss term corresponds to the feedback mechanism of discriminator network.

Song et al. [88] proposed a method for efficient architecture search for SR networks. The proposed method utilizes a RDB as key block and employs a search algorithm to explore the space of RDB configurations. The search is guided by a proxy task of LR image classification, which helps identify optimal configurations of RDBs for image SR. The authors have presented their findings on a resulting network, named RDN-E, which exhibits state-of-the-art (SOTA) performance on various benchmark datasets while requiring significantly fewer parameters and computational resources than existing approaches.

Although GAN based SR has produced impressive results in photorealistic images, its high memory consumption has prevented its widespread use in mobile devices with limited resources. To solve this problem, [89] introduces a PatchGAN discriminator for quicker and more frequent training and uses a memory-efficient network as the generator. Both the compressed generator and the student discriminator are distilled to achieve equilibrium. A hardware-aware neural architecture search is also carried out to locate a specific sub-generator for the intended mobile device. This framework makes GAN based SR more appropriate for portable devices by lowering the memory access cost of the generator while retaining performance.

The instability of GANs makes training them difficult because it frequently produces undesirable artifacts in addition to the necessary features. To solve this issue, [90] presented a technique that uses local statistics like residual variance to distinguish between GAN generated artifacts and authentic details. The proposed method, locally discriminative learning (LDL),

produces a more perceptually accurate and artifact-free SR image by creating an artifact map to govern and stabilize the model training process.

Images produced by SR models using GANs frequently have structural flaws. These problems have been addressed by a novel approach called high-frequency information fusion GAN [91]. High-frequency information that is crucial to the human eye is incorporated into the SR process in this approach. It accomplishes this by enhancing the ESRGAN model's network architecture and constructing a compact spatial attention module to extract high-frequency data.

3.1.5 Attention based networks

Residual Channel Attention Networks (RCAN) is a CNN approach for SR, introduced by Zhang et al. [92]. The architecture comprises a series of residual channel attention blocks (RCABs) interconnected via skip connections. The RCABs are specifically devised to enhance the network's capacity for representation by dynamically scaling the importance of features. The skip connections facilitate information propagation between different hierarchical levels within the network.

Three main issues are addressed by Zhang et al. [93] on the existing CNN methods: small receptive field, unable to distinguish between the plain area and textural area, and low and high frequency components are treated equally. To overcome these issues, they proposed a RNAN model where residual attention blocks based on local and non-local areas are combined to obtain information of hierarchical features to generate a SR image.

Dai et al. [94] introduced a module based on channel attention to modulate the significance of individual channels within a feature map. This module enables the network to concentrate on the most informative features and enhance its discriminative capabilities for object differentiation. Additionally, the residual connection based structure integrates non-local operations to effectively capture long-range dependencies. This structural component facilitates the network's acquisition of pixel relationships that span considerable distances, which holds substantial importance in the context of SISR.

Yan et al. [95] proposed a lightweight multi-scale spatial attention networks (MSAN) SR network. MSAN comprises a feature extraction stage, an MSA module, and a reconstruction stage. The MSA module is designed to learn spatial dependencies between different feature maps at multiple scales and uses a pyramid-like structure to capture both local and hierarchical features before the final reconstruction stage.

[96] proposed dual-view attention network consists of two sub-networks, one focusing on image details and the other on global features. The two sub-networks are combined through a dual-view attention mechanism that utilizes both spatial and channel attention maps to highlight

important features. The attention maps are used to re-weight the feature maps before they are merged.

The existing approaches for channel attention have demonstrated effectiveness in preserving features rich in information within individual layers of deep learning models. However, these approaches often treat each convolution layer as an independent entity, which may lead to sub-optimal performance. To overcome this limitation, a novel holistic attention network (HAN) is proposed by [97] that captures comprehensive interdependencies among layers, channels, and positions. The HAN architecture comprises two attention modules: a layer attention module (LAM) and a channel-spatial attention module (CSAM).The LAM intelligently highlights hierarchical features by incorporating correlations among layers, while the CSAM selectively captures highly informative features by learning positional confidence across all channels.

[98] proposed the multi-path residual network (MPRNet) comprises three key components: 1) Multi-path residual block containing a set of parallel residual paths with different dilation rates to capture different scales of information; 2) Depthwise separable convolution that reduces the computational cost and model complexity; and 3) Channel attention module that enhances the feature representations by re-calibrating channel-wise feature responses.

Lu et al. [99] presented an approach to enhance the network's capability of capturing essential features in SR context. Their proposed method involved the integration of a convolutional block attention module (CBAM) within a dense block, facilitating the efficient exchange of information across feature maps. Moreover, the researchers introduced a spatial module that leveraged the self-attention mechanism to effectively capture long-range spatial connections, ensuring stability in the feature extraction process.

Behjati et al. [100] presented a novel network architecture for SISR that effectively balances computational efficiency and accuracy. Their approach incorporates a directional variance attention (DiVA) mechanism, enabling the concurrent capture of long-range spatial dependencies and inter-channel relationships, thereby producing more informative representations. Additionally, the authors introduced a residual feature group to parallelize the computation of attention and residual blocks. By linearly fusing the outputs of each residual block at the RAFG output, the network gains access to the comprehensive feature hierarchy, enhancing its overall performance.

3.1.6 Frequency based networks

Discrete cosine transform (DCT) based methods employ a representation strategy that characterizes feature maps in the frequency domain instead of the spatial domain

For various applications, CNNs are being trained in frequency domain, such as, Zhang et al. [101] extended the idea of DCT coefficients and presented median filtering forensics approach

which was based on a CNN with an adaptive filtering layer (AFL) built in the DCT domain. Meanwhile, Verma et al. [102] addressed the problem of classifying images based on the number of JPEG compressions they have undergone, by utilizing deep CNNs in DCT domain.

For SR task, Islam et al. [103] used directional fourier phase feature components to adaptively learn the regression kernel based on local co-variance to estimate the SR image.

Li et al. [104] introduced an innovative neural network for image SR that operates in the frequency domain. The network capitalizes on the convolution theorem to transform spatial domain convolutions into frequency domain products. Furthermore, the conventional non-linearity achieved through rectifier units in deep networks is effectively realized as a frequency domain convolution. This strategy guarantees computational efficiency during testing while facilitating parameter learning through backpropagation. The network employs the Hartley transform as an alternative to the Fourier transforms, eliminating the requirement for complex numbers.

Guo et al. [105] proposed a DCT-DSR network that utilizes a convolutional DCT (CDCT) layer to integrate DCT into the network structure. The CDCT layer is further extended to become trainable by imposing orthogonality constraints on the individual basis functions. The proposed orthogonally regularized deep SR network (ORDSR) takes advantage of the image transform domain while adapting the design of the transform basis to the training image set.

3.1.7 Cross-modality support based networks

Semantic correlation between the audio and visual information is utilized in numerous computer vision problems [106, 107]. For example, Tian et al. [108] used the aural information for the comprehensive study of scene.

[109, 110] proposed an attention mechanism between the cross-modal aural-visual network to abolish the temporal inconsistency during localization of events and analyze the longer videos with prominent information, respectively. By exploring the relation between speech and visual representation, Wen et al. [111] and Oh et al. [112] proposed GAN networks that generates facial images from the speech using physical attributes like identity matching, age and gender etc.

Chen et al. [113] converted the aural signals into a complex structure which corresponds to facial landmarks and using these landmarks facial images are generated. Zhang et al. [114] embedded the aural information in the CNN architecture to enhance the facial videos and remove compression deformities.

Along with LR image encoder, the audio encoder is used by [115] for image SR. Feature maps obtained by amalgamating the hierarchical features of both the encoders are applied to the decoder, resulting in a high resolution image.

3.2 Facial image SR

Depending upon the information utilized by the face super resolution (FSR) methods to generate an output image, CNN based methods are broadly classified into three main categories. 1) global methods (an entire face image is fed as input for SR network), 2) local methods (facial components are used in the network to obtain a high resolution image) and 3) mixed methods.

3.2.1 Facial image SR by CNN

Huang et al. [116] introduced a method for obtaining HR face images by utilizing a multiple filter convolution technique for feature map extraction, followed by a non-linear mapping. Liu et al. [117] introduced a novel approach that incorporates an iterative back projection technique as a post-processing technique within a CNN based framework designed for FSR. The CNN is initially trained to generate a SR image from a LR input image. However, despite the initial SR image produced by the CNN, it may still exhibit inherent imperfections and artifacts. In order to address this concern, an iterative back projection method is employed to iteratively refine the resulting image, leveraging the initial output of the CNN as a starting point for enhancement.

Attention based networks have achieved great performance in general SR [92]. Taking inspiration from these networks, spatial attention is introduced by chen et al. [118] and channel attention mechanism by Chudasama et al. [119] in CNNs, guiding the FSR system to generate output images with sharp key features.

To add textural information in the FSR network, wavelet based network are proposed by [120, 121]. The proposed architecture is presented to effectively predict SR wavelet coefficients, thereby achieving improved sharpness in facial images. By incorporating prior knowledge of facial characteristics, the model selectively accentuates significant facial features. Moreover, the network integrates a linear low-rank convolution methodology to optimize the SR process further.

Different facial parts are generated using different strategies in local methods. Hu et al. [122] introduced a FSR method where face image is decomposed into high frequency enhanced face and low frequency basic face using sparse representation and deep convolution networks, respectively. The output obtained from the two networks is then fused to get the final image.

Feng el al. [123] proposed a SR network where the outputs obtained from two patch based auto encoders is merged with traversal network to generate final SR image. To incorporate both local and global details within the face hallucination network, Lu et al. [124] introduced a fused network that combines global and local information. This network adds high frequency details by initially learning information at local level and then gradually moving to global level in the architecture.

3.2.2 Facial image SR using GAN

Earlier works in the field of FSR using GANs include numerous renowned works, some of which are mentioned below:

Enhancing facial images to ultra-resolution using a discriminator network is a notable approach [125]. In this framework, the discriminator receives both the synthesized image from the generator network and the ground truth HR image, compelling the generator to mimic HR images.

To address the challenges associated with training GANs and ensure training stability, Chen et al. [126] introduced utilizing Wasserstein distance as a training metric in FSR. Furthermore, Huang et al. [127] proposed a SR network based on GANs, where the generator and discriminator components consist of AEs. The network is optimized using a combination of GAN loss and pixel-wise loss to achieve improved results.

Inception architecture based GAN is presented by Indradi et al. [128] to obtain face images with HR. Luo et al. [129] proposed a SR network, where an upsampling module composed of encoder, decoder and upsampling layers. This module is designed to incorporate prior facial information, and it is combined with a discriminator network to optimize the SR process.

In place of traditional discriminator, Zhang et al. [130] proposed a pixel-wise discriminator, designed to receive two input arguments. First input is either the ground-truth image or the generated image and the second input comprises facial characteristics obtained from a pre-existing facial analysis model, prompting the generator to generate images with enhanced texture and finer details.

Cheng et al. [131] combined the traditional SR method with unsupervised domain methods by commencing characteristic regularization between two CNNs. This technique improves the gradient flow between the networks and hence network is able to generate images with HR. Ko et al. [132] argued that recent GAN based methods require extra information along with the LR image to generate images with fine perceptual details. But they utilized only a LR image with its edge information at various scales to generate HR image.

Most GAN based methods use bicubic kernels to obtain LR images from the HR image for training. So, the training dataset does not follow the natural degradation process, which affects GAN-based method's performance on realistic LR images. To address this problem, Aakerberg et al. [133] introduced different types of noises in LR images for training dataset.

3.3 Video super resolution

Enhancing the resolution of videos, commonly referred to as video super-resolution (VSR), aims to improve the quality and level of detail in the content. Deep learning techniques have

been increasingly popular in VSR as they excel at acquiring intricate representations by leveraging large datasets.

The VSRnet2 [134], a VSR approach, was built on the SRCNN [67] method. Its network architecture constitutes a module for motion estimation along with the convolution layer. In contrast to SRCNN, which uses a single LR image as input, VSRnet2 makes use of a series of motion-corrected frames that are taken one after the other to achieve HR.

Providing high-quality content for ultra-high-definition televisions requires VSR. The majority of deep learning-based methods depend on precise motion estimation and compensation. Jo et al. [135] instead suggested a new approach that creates dynamic upsampling filters and a residual image depending on the immediate spatiotemporal neighbourhood of each pixel using a DNN.

[136, 137] presented a VSR method that used a combination of 3D convolution and bi-directional LSTM to map temporal dependency across adjacent frames and long-term dependency across all the frames of videos simultaneously. However, 3D convolutions have high computational complexity, which makes this network training challenging.

To capture temporal dependencies across frames, a novel feature extraction module was introduced by Haris et al. [138]. This module serves a dual purpose: extracting feature maps from the input frame and aligning the concatenated feature maps derived from the input and neighboring frames. Subsequently, an encoder-decoder architecture based projection module is employed to extract hierarchical features necessary for achieving the desired resolution.

To map the temporal information across the frames, the Haris et al. [138] proposed a feature extraction module with two functionalities: extracting the feature maps from the input frame and alignment performed on the feature maps extracted after concatenation of the input and neighboring frames. This module is followed by a projection module based on the encoder-decoder architecture to extract the hierarchical features to achieve the target resolution.

Zhu et al. [139] divided the VSR problem into three domains: firstly, by using inverse residual block in parallel, authors tried to map the spatial information. The extracted spatial feature maps are then applied to the ConvLSTM network to map the temporal features. Finally, the adaptive sparse fusion technique selects the crucial features and generates a HR video.

Since L2 losses oversimplify the spatial details leading to smooth results across the generated video, Chu et al. [140] presented a self-supervised learning approach, where temporal coherence is attained without losing the spatial information using temporal adversarial learning.

Optical flow for the temporal alignment introduce artifacts in the generated video. Hence adaptive spatial filters are proposed by Wen et al. [141] for VSR. Building blocks for the proposed architecture are residual blocks in combination with the channel attention layer.

Chan et al. [142] empirically discovered that using a longer sequence of frames rather than a bigger batch size is more effective in retaining temporal consistency while training a VSR network. They also introduced a variety of degradations into the LR input frames so that the proposed algorithm can generalize well for the real-world LR videos.

Spatiotemporal information is investigated more comprehensively for the alignment of adjacent frames across the videos by Chan et al. [143] They used a recurrent structure in conjunction with second-order grid propagation to generate high-resolution output videos.

3.4 Summary

Deep neural networks have exhibited significant advancements in the domain of high-resolution image generation. Notably, deep learning models, particularly those employing Generative Adversarial Networks (GANs), have proven to be efficacious tools for addressing super-resolution challenges. These GAN-based architectures excel at rendering images with remarkable fidelity, precise details, and realistic textures, rendering them virtually indistinguishable from their ground-truth counterparts.

However, earlier GAN-based methodologies encountered certain impediments, such as the computational complexity associated with training and the issue of vanishing gradients. Moreover, these methods were predominantly reliant on 2D information, potentially leaving out crucial 3D aspects, like depth and structural information, necessary for generating truly comprehensive high-resolution images. Additionally, in the context of video super-resolution, temporal inconsistency issues plagued the quality of the generated frames.

Hence, within the scope of this dissertation, our objective was to rectify the limitations observed in prior research. We present a comprehensive exposition of our novel frameworks, meticulously designed to overcome the intricacies inherent in the super-resolution problem, thereby addressing the aforementioned challenges.

4 Semantic Information Based Image Super-Resolution System

4.1 Introduction

The common measure to optimize the SR algorithm is mean square error (MSE), and its performance is evaluated using peak signal-to-noise ratio (PSNR) [144]. In PSNR based approaches, error is minimized in the pixel space instead of the feature space. So, these approaches tend to measure perceptual results poorly due to their inefficacy to measure high frequency components. Therefore, GAN with adversarial loss function [69] are proposed to obtain visually pleasing results. In these networks, loss functions data is minimized in the feature space and visual results are calculated using mean opinion score testing method. Thus, they are able to produce SR results that are very similar to the real images [145], as shown in Figure 4.1.

Our main contributions are as follows:

- Two architectures have been proposed. First architecture is Residue based Dual Subpixel Generative Adversarial Network (RDS-GAN). This model consists of two stages: Premier Residual Stage (PRS) and Deuxieme Residual Stage (DRS). These two stages of subpixel layer are used to increase resolution of an image. Two stage upsampling process provides better learning capability to our model, allowing it to capture fine texture details of an image. Dense blocks are connected within these two stages, where inter and intra residual connections are made to preserve the high-frequency details.
- The second proposed architecture is Semantic feature based Dual Subpixel Generative Adversarial Network (RSDS-GAN). In this model, we have introduced object based enhancement in the generator architecture. Feature maps extracted from VGG19 pre-trained model are merged with the input image to embed semantic information in generator.
- Spectral normalization is introduced in the discriminator architecture to stabilize the training process for SR.

The architectures proposed in this chapter outperforms previous SOTA methods and produce visually more appealing solutions.

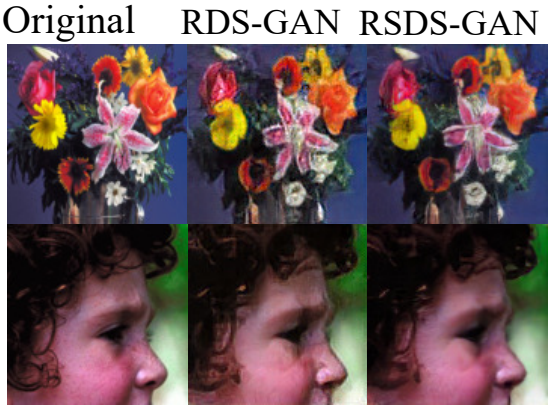


Figure 4.1: Output image (right) is almost same as the real image (left).

4.2 Proposed Architecture

Our primary aim to train the Generator Network (G_{ξ_g}) with the help of discriminator is to get a HR image from its LR counterpart. To achieve this aim, a novel architecture for generator is proposed as illustrated in Figure 4.2. We divide generator network into two stages: Premier Residual Stage (PRS) and Deuxieme Residual Stage (DRS). These two stages provide better feature learning capability to our model (discussed in section (4.3.1)).

As shown in Figure 4.2 , a low-resolution image (i^{lr}) and feature maps (i^{f_m}) produced from VGG19 pre-trained model are passed to PRS as an input to convolution layer producing F^1 and F^2 (refer to eq. (4.1),(4.3)) on which LeakyReLU activation function (Θ) is applied resulting in $L^1$and L^2 (eq.(4.2),(4.4)).

$$F^1(i^{lr}; w_1, b_1) = w_1 \otimes i^{lr} + b_1 \tag{4.1}$$

$$L^1 = \Theta(F^1(i^{lr}; w_1, b_1)) \tag{4.2}$$

$$F^2(i^{f_m}; w_2, b_2) = w_2 \otimes i^{f_m} + b_2 \tag{4.3}$$

$$L^2 = \Theta(F^2(i^{f_m}; w_2, b_2)) \tag{4.4}$$

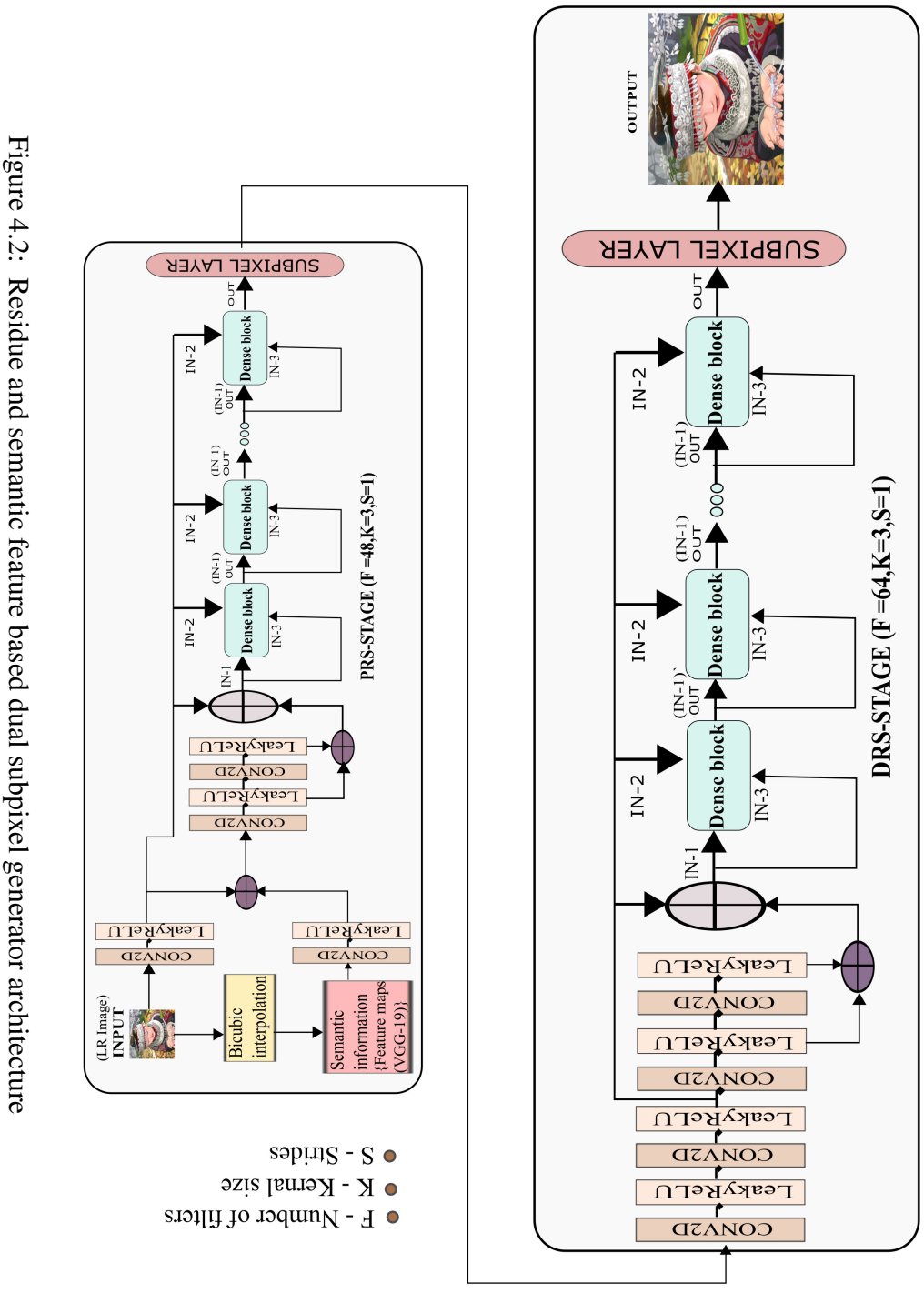


Figure 4.2: Residue and semantic feature based dual subpixel generator architecture

To embed semantic based information in the generator $L^1$and L^2 are merged resulting is eq. (4.5).

$$L^3 = L^1 + L^2 \tag{4.5}$$

L^3 is then passed to following convolution layers activated with LeakyReLU activation function producing F^k and L^k (refer to eqs. (4.6,4.7) where $k \, \epsilon \, [4, 5]$)

$$F^k(L^{k-1}; w_k, b_k) = w_k \otimes L^{k-1} + b_k \tag{4.6}$$

$$L^k = \Theta(F^k(L^{k-1}; w_k, b_k)) \tag{4.7}$$

High frequency feature maps are created by residual learning. First residue ($\Re^1$) is generated by merging L^4 and L^5, obtained from the eq. (4.7) . Further $\Re^1$ from eq. (4.8) is merged with L^1, generated from eq. (4.7), to create second residue ($\Re^2$).

$$\Re^1 = L^4 + L^5 \tag{4.8}$$

$$\Re^2 = \Re^1 + L^1 \tag{4.9}$$

The output $\Re^2$ is passed to dense block (refer to Figure (4.3)) arranged sequentially with depth of 8, refer to eqs. $(4.14 - 4.19)$

$$F^v(\Re^{u-1}; w_v, b_v) = w_v \otimes \Re^{u-1} + b_v \tag{4.10}$$

$$L^v = \Theta(F^v(\Re^{u-1}; w_v, b_v)) \tag{4.11}$$

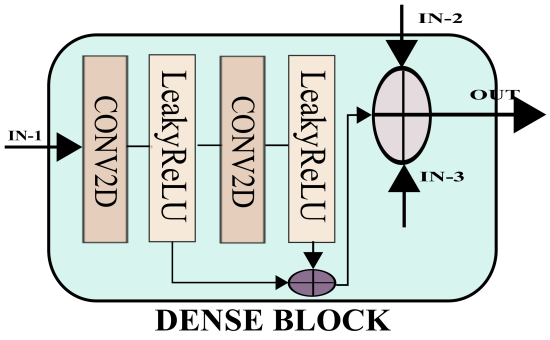


Figure 4.3: Dense block architecture

$$F^{v+1}(L^v; w_{v+1}, b_{v+1}) = w_{v+1} \otimes L^v + b_{v+1} \tag{4.12}$$

$$L^{v+1} = \Theta(F^{v+1}(L^v; w_{v+1}, b_{v+1})) \tag{4.13}$$

$$R^u = L^v + L^{v+1} \tag{4.14}$$

$$\Re^{u+1} = \Re^u + \Re^{u-1} + L^1 \tag{4.15}$$

Here $u\epsilon[3, 10]$ and $u\epsilon[6, 13]$

The output from 8th dense block is fed to the first subpixel layer as shown below:

$$S^1(L^{14}; w_{15}, b_{15}) = \phi\left(w_{15} \otimes \Re^{10} + b_{15}\right) \tag{4.16}$$

$\phi(.)$ represents ReLU activation function, and $\Re^{10}$ is the output of 8th dense block. So, the input image is upscaled with a factor of two by using subpixel layer.

The output feature maps, obtained from PRS stage, are passed to the DRS stage. Operations performed in DRS stage are same as in PRS stage, except the feature map size and the operation performed in eqs. (4.3) and (4.4). The feature maps in this stage are of double size than the

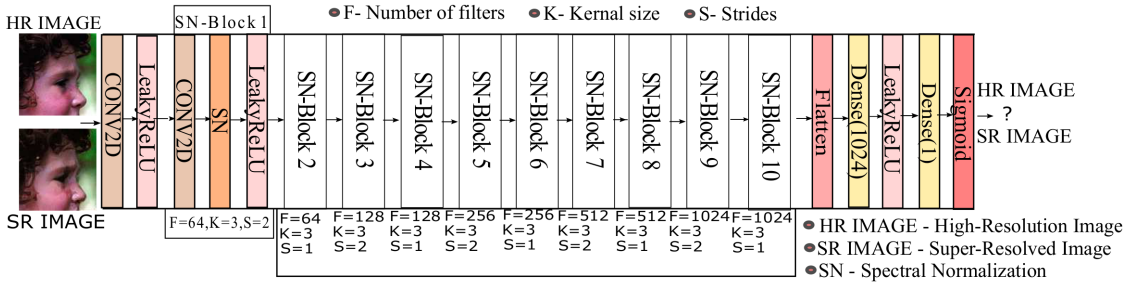


Figure 4.4: Discriminator architecture

previous stage. From the second subpixel layer we get the final image ($I^{\mathcal{F}}$) which is four times of the ground-truth image.

A Discriminator Network, denoted as D_{ξ_d}, is trained to discern authentic images from those generated by the generator network. The primary objective of training the generator network is to deceive the discriminator network, enabling the generator to produce outputs resembling ground-truth images. The architecture of the discriminator network is depicted in Figure (4.4). In order to enhance the discriminator's performance, we have incorporated spectral Normalization, a technique introduced by Ledig et al. [69]. Spectral Normalization, proposed by Miyato et al. [146], restricts the discriminator's Lipschitz constant. This constraint facilitates a more balanced training process within the GAN framework. Notably, spectral Normalization offers the advantage of minimal computational complexity, as it does not necessitate the tuning of additional hyperparameters.

4.2.1 Loss function

Generator network G_{ξ_g} is dependent on parameters ξ_g. ξ_g refers to two parameters: weights and biases $(w_{1:v}; b_{1:v})$, v refers to the number of layers used in generator architecture. To obtain final output image $I^{\mathcal{F}}$, the loss function (L^f) is optimized over M training samples and the parameters (weights and biases) are updated according to the optimization technique represented in Eq. (4.17):

$$\tilde{\xi_g} = \underset{\xi_g}{argmin} \frac{1}{M} \sum_{m=1}^{M} L^f(G_{\xi_g}(i_m^{lr}), i_m^{hr}) \tag{4.17}$$

Here, i_m^{hr} are the HR images used for training, and i_m^{lr} are the corresponding LR images. Loss function (L^f) is formed by combining losses discussed in eqs. (4.18, 4.19 and 4.20).

To obtain a SR image having perceptually suitable features, a loss function L^f has been defined based on Ledig et al. [69] and Bruna et al. [83] instead of using MSE. This loss function combines two distinct components: the adversarial loss and the content loss.

$$L^f = L^f_{vgg} + L^f_{gen} \tag{4.18}$$

Here, L^f_{vgg} is a VGG features based content loss. Instead of using pixel based MSE loss, perceptual loss [147] is calculated. It satisfies the perceptual quality of an image. Euclidean distance between the feature maps generated by VGG-19 (pre-trained model) from generated image $G_{\xi_g}(i^{lr})$ and the ground-truth image (i^{hr}) is calculated to find similar feature representation constitutes the VGG loss (refer to eq. 4.19)

$$L^f_{vgg/i.j} = \frac{1}{W_{i,j}H_{i,j}} \sum_{a=1}^{W_{i,j}H_{i,j}} \sum_{b=1} (\varphi_{i,j}(i^{hr})_{a,b} - \varphi_{i,j}(G_{\xi_g}(i^{lr}))_{a,b})^2 \tag{4.19}$$

Here, feature maps are represented by $\varphi_{i,j}$, which are obtained from the VGG network by extracting the features from $j-th$ convolution layer prior to the $i-th$ maxpooling layer. Feature map dimensions of VGG network are represented by $W_{i,j}$ and $H_{i,j}$.

The second loss component of L^f is adversarial loss. This loss component lets our architecture to use generative element for producing SR images that look similar to the natural images by fooling a discriminator. This loss function relies on the probabilities given by the discriminator on the generator's output as shown below:

$$L^f_{gen} = \sum_{m=1}^{M} -logD_{\xi_d}(G_{\xi_d}(i^{lr})) \tag{4.20}$$

Here, $D_{\xi_d}(G_{\xi_d}(i^{lr}))$ shows the probability that the output image generated from the generator $(G_{\xi_d}(i^{lr})$ is a ground-truth image.

4.3 Experiments, results and discussions

We have evaluated our model on publicly available benchmark datasets. First, analysis is done for subpixel layer, then a brief introduction is given about the dataset used for training and testing our architecture. After that a comparative analysis is presented between our model and other SOTA techniques.

4.3.1 Analysis for subpixel layer

We analyzed the effect of using subpixel layer at different positions in the architecture. Three cases are considered to evaluate the performance of subpixel layer. In the first case, as shown

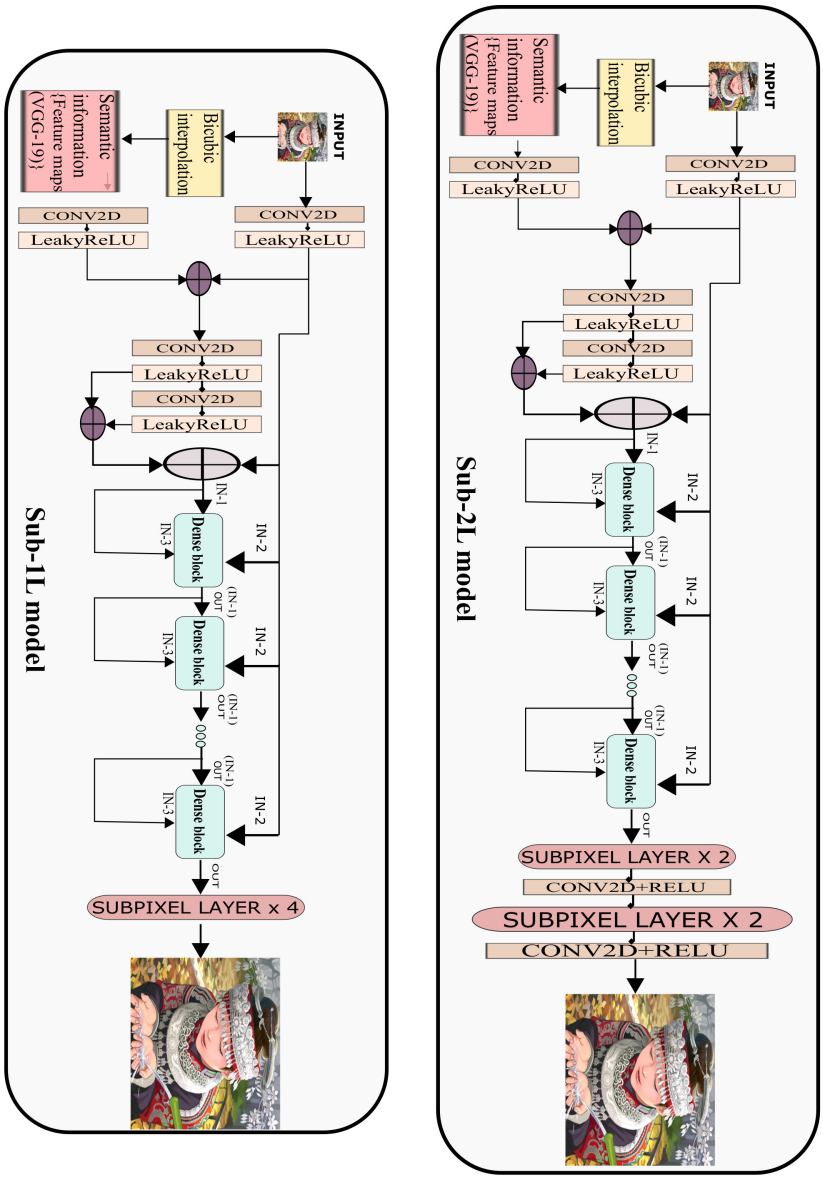


Figure 4.5: Generator model with one subpixel layer and 2 subpixel layers respectively

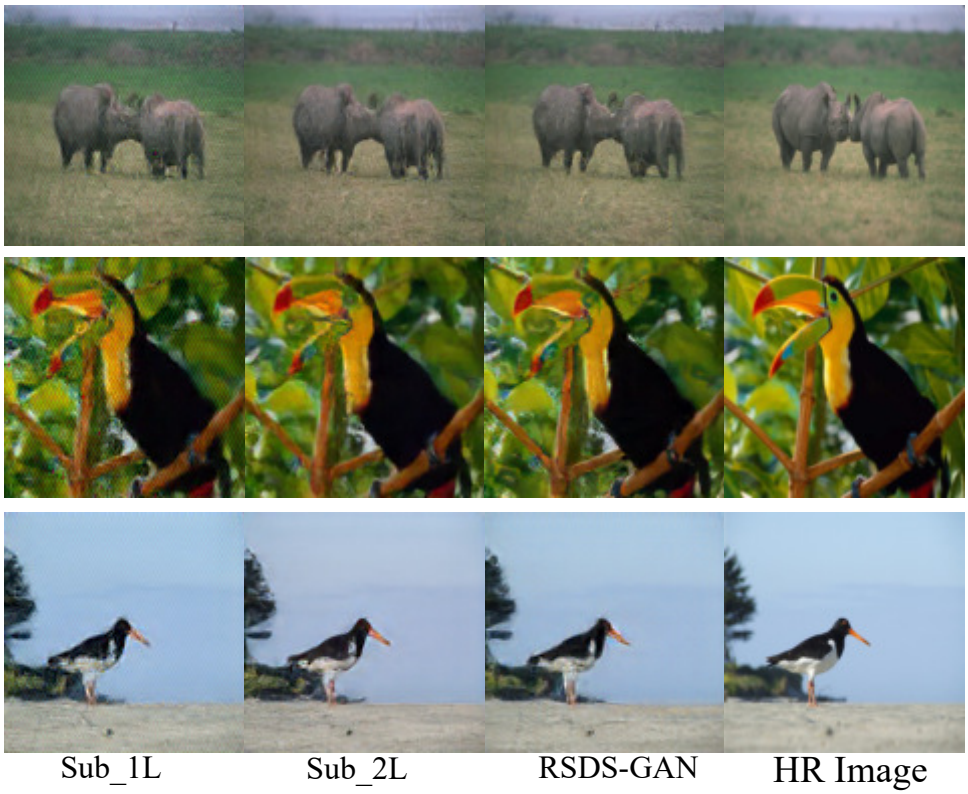


Figure 4.6: Perceptual results showing effects of using subpixel layer at different positions.

Table 4.1: Performance comparison using quantitative values for different test datasets to show the effect of using subpixel layer at different positions.

Models	Sub_1L	Sub_2L	RSDS-GAN (ours)
Metric	MSE/PSNR(dB)	MSE/PSNR(dB)	MSE/PSNR(dB)
BSD200	0.0031/25.02	0.0024/26.06	0.0021/26.67
SET14	0.0071/21.47	0.0054/22.63	0.0056/22.46
Imagenet	0.0037/24.21	0.0038/24.18	0.0030/25.02

in Figure (4.5), a subpixel layer is used as the last layer of the architecture (Sub_1L model) with an upscaling factor of 4. Visual and quantitative results obtained from this model are poor. This model is not capable to learn the fine details present in an image. In second case, we use two subpixel layers (Sub_2L model) instead of a single subpixel layer as shown in Figure (4.5). Results obtained from this model are perceptually good with high quantitative values as compare to the first case. Third case is the proposed model (shown in Figure (4.2)). In this case, upscaling is done in two stages. By using two stage enhancement process, our model is able to mimic each minor detail of an image and provide results very similar to the ground-truth image. Perceptual results from these models on test datasets (SET5 [148], BSD200 [149], SET14 [150] and imagenet [151] are shown in Figure (4.6) and corresponding quantitative analysis is shown in table (4.1). From the qualitative and quantitative values, it is clear that the two stage subpixel layer model (RSDS-GAN) provides superior results than the Sub_1L and Sub_2L model.

4.3.2 Datasets

To assess the performance of the proposed architecture, an upscaling factor of 4 was employed, indicating a four-fold increase in size between the input and output images. The dataset utilized for evaluation consisted of a collection of $50,000$ HR images. To generate LR counterparts for the HR images, a downsampling process was applied using a bicubic kernel. Specifically, the HR images, sized 120×120, were downsampled by a factor of 4, resulting in low-resolution images of size 30×30. In order to incorporate semantic information, feature maps were extracted from the ninth layer of a pre-trained VGG19 model. These extracted features, with dimensions of 30×30, were obtained by feeding bicubic versions of the low-resolution images, which were sized at 120×120, into the VGG19 model. The training process utilized a dataset composed solely of high-resolution images. The Adam optimizer [152] was employed with specific hyperparameters: a learning rate of 0.0001, β_1 set to 0.9, and β_2 set to 0.999. The loss metric L^f (defined in equation (4.18)) was utilized to quantify the loss during training. The model was trained for 40 epochs, with training performed in batches using randomly sampled examples from the training dataset, each batch consisting of 32 samples. The generator and discriminator models were updated alternately until the model reached convergence. Two commonly adopted evaluation metrics were employed for quantitative analysis of the image super-resolution techniques: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measurement (SSIM). These metrics are widely used to assess the quantitative results of SR tasks and provide valuable insights into the fidelity and similarity of the enhanced images compared to the ground truth.

4.3.3 Comparison with SOTA methods

We evaluate our models performance with other SOTA methods: SRCNN [153], ESPCNN [68], SRGAN [69], ESRGAN [84] and bicubic interpolation. All these models are trained on imagenet training dataset for 40 epochs for fair comparison. The quantitative analysis is presented in table 4.2 shows the average values obtained on SOTA methods and our models. Visual results are shown in Figures 4.8, 4.9, 4.10, 4.11, 4.12, 4.13 on the globally used benchmark datasets (SET5, SET14, BSD200) and the test dataset generated from the imagenet dataset with their quantative analysis.

Variation of generative loss across epochs is shown in Figure (4.7) for SRGAN [153] and RDS-GAN and RSDS-GAN. It is clear from the comparison that our method RDS-GAN is achieving lower error rate than SRGAN but RSDS-GAN convergence error rate is more than the other two methods. Based on an evaluation of previous GAN-based SR methods, such as those proposed by Ledig et al. [69] and Wang et al. [84], as well as our own comprehensive visual and quantitative analysis, it becomes evident that high quantitative metrics values do not consis-

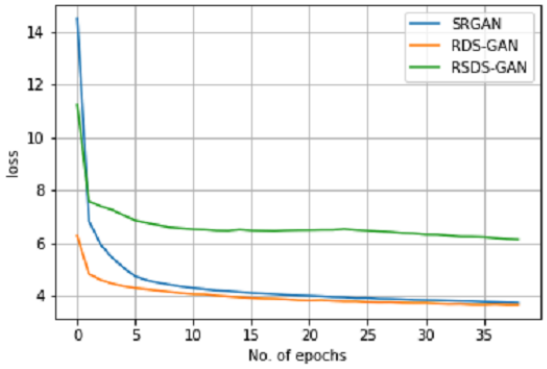


Figure 4.7: Comparison graph of generator loss against number of epochs for SRGAN and RDS-GAN

tently correlate with superior perceptual results. Regarding visual quality, our proposed method surpasses all other SOTA techniques, as our rigorous assessment supports.

4.3.4 Mean Opinion Score (MOS) evaluation

As introduced by Ledig et al. [69], a MOS test was conducted to assess the visual quality of the generated output images. This test involved the evaluation of outputs obtained from various SOTA methods, including SRCNN [153], ESPCNN [68], SRGAN [69], ESRGAN [84], bicubic interpolation, and our proposed methods, namely RDS-GAN and RSDS-GAN. The MOS test results, presented in Figures 4.8, 4.9, 4.10, 4.11, 4.12, and 4.13, were obtained by soliciting ratings from 16 expert raters. Raters were asked to assign integral scores ranging from 1 (indicating poor quality) to 5 (reflecting excellent quality) to randomly selected output images from datasets such as SET5, SET14, BSD200, and ImageNet. The average values were then calculated based on the integral scores provided by the raters. The results in the figures demonstrate that our proposed method exhibits superior performance compared to other SOTA methods in terms of visual quality, as verified by the MOS test.

4.4 Conclusion

A novel GAN based architecture is proposed for SR where dual stage upsampling is done for an upscaling factor of 4 in this chapter. In dual upsamping stages, inter and intra residual dense connections are done; making our model capable of sustaining high texture details of an image. To enhance the objects present in the image, semantic information is merged with the input image; leading to excellent visual results. To stabilize the training process spectral normalization is used in the discriminator architecture. Qualitative and quantitative results are

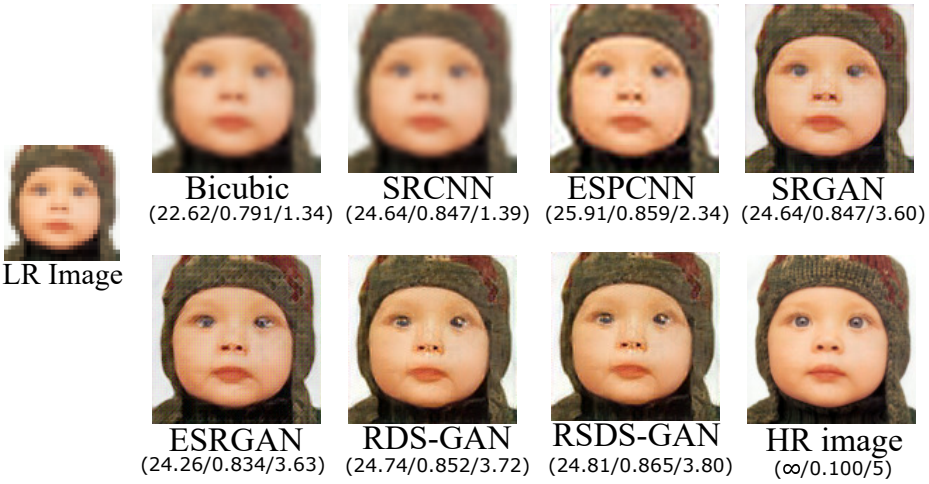


Figure 4.8: The perceptual results, accompanied by their corresponding PSNR/SSIM/MOS scores, on the 'baby' image (SET5) with the upscaling factor 4 using various SOTA algorithms.

Table 4.2: Performance comparison on the basis of average MSE, average PSNR (dB) and average SSIM for various SR methods on various test datasets (SET5, SET14, BSD200 and Imagenet) with scale factor of 4.

Dataset	SET5			SET14			BSD200			Imagenet test dataset		
Metric	MSE	PSNR	SSIM	MSE	PSNR	SSIM	MSE	PSNR	SSIM	MSE	PSNR	SSIM
Bicubic	0.0094	21.65	0.720	0.0083	21.34	0.686	0.0059	22.86	0.722	0.0057	23.21	0.766
SRCNN [153]	0.0077	22.83	0.701	0.0100	21.24	0.706	0.0089	21.04	0.677	0.0063	22.67	0.755
ESPCNN [68]	0.0056	23.82	0.811	0.0058	22.93	0.764	0.0040	24.57	0.795	0.0035	25.30	0.833
SRGAN [69]	0.0066	22.99	0.799	0.0064	22.50	0.751	0.0051	23.44	0.762	0.0047	23.93	0.801
ESRGAN [84]	0.0058	23.33	0.794	0.0060	22.78	0.749	0.0052	22.77	0.702	0.0048	23.70	0.786
RDS-GAN (ours)	0.0067	22.94	0.804	0.0072	22.16	0.746	0.0060	22.73	0.746	0.0053	23.42	0.789
RSDS-GAN (ours)	0.0062	23.13	0.832	0.0072	22.12	0.762	0.0056	23.02	0.770	0.0050	23.64	0.812

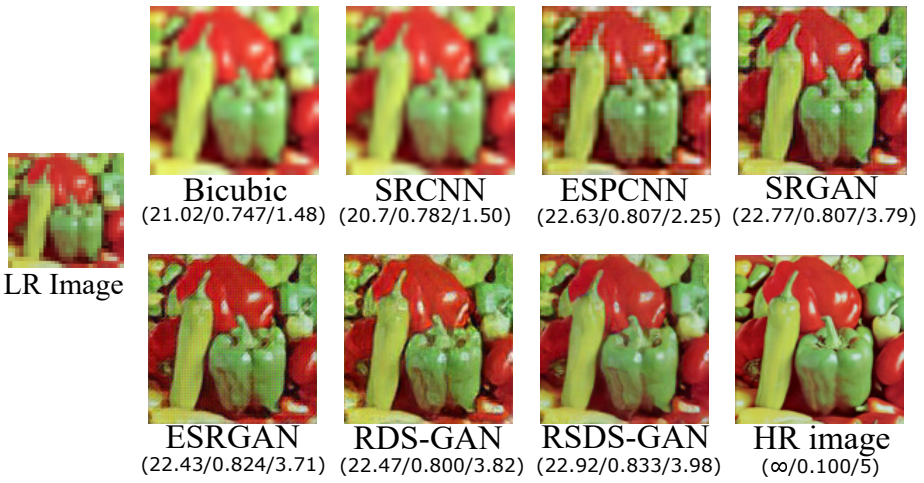


Figure 4.9: The perceptual results, accompanied by their corresponding PSNR/SSIM/MOS scores, on the 'pepper' image (SET14) with the upscaling factor 4 using various SOTA algorithms.

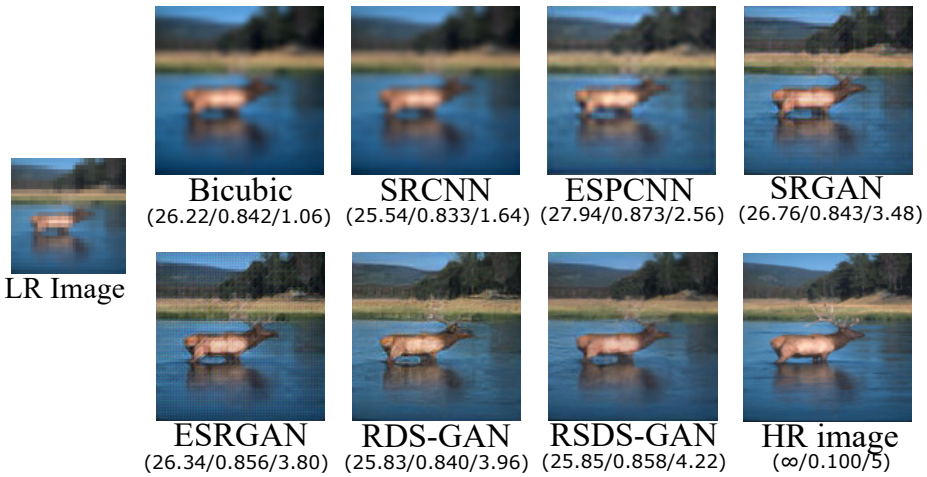


Figure 4.10: The perceptual results, accompanied by their corresponding PSNR/SSIM/MOS scores, on the '104055' image (BSD200) with the upscaling factor 4 using various SOTA algorithms.

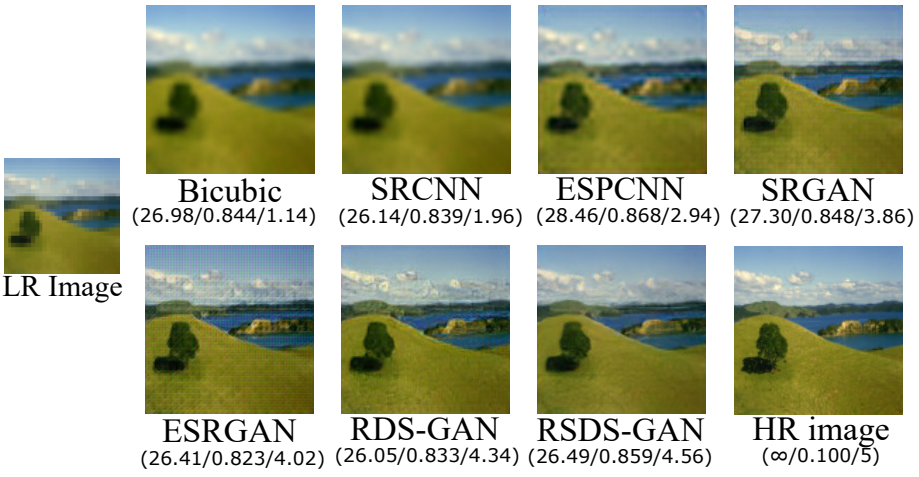


Figure 4.11: The perceptual results, accompanied by their corresponding PSNR/SSIM/MOS scores, on the '36046' image (BSD200) with the upscaling factor 4 using various SOTA algorithms.

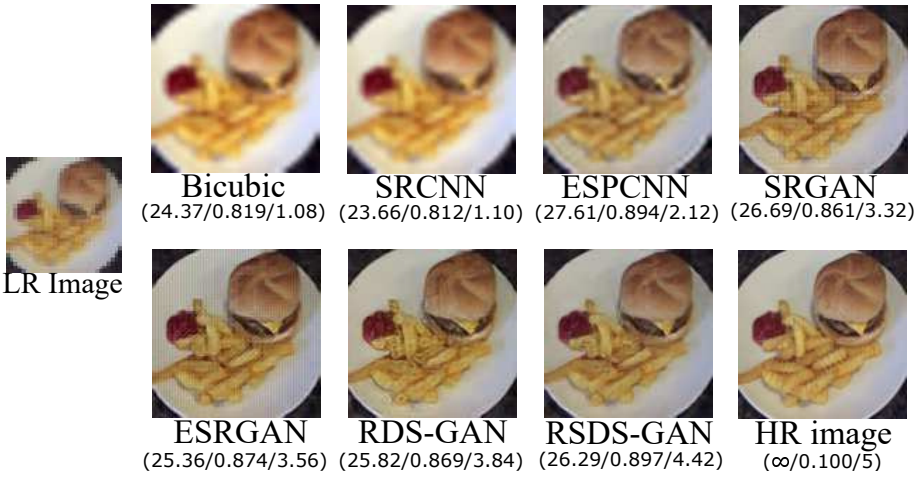


Figure 4.12: The perceptual results, accompanied by their corresponding PSNR/SSIM/MOS scores, on the test image (imagenet) with the upscaling factor 4 using various SOTA algorithms.

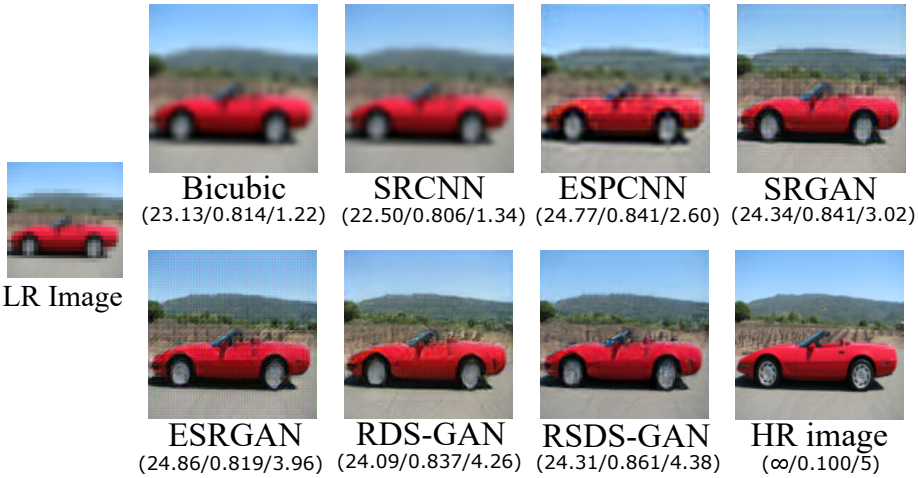


Figure 4.13: The perceptual results, accompanied by their corresponding PSNR/SSIM/MOS scores, on the test image (imagenet) with the upscaling factor 4 using various SOTA algorithms.

calculated for our method and other SOTA methods. Comparison is done using MOS values to verify the perceptual quality of images produced by our method against other SOTA methods. From the visual results and MOS testing values it is clear that our method outperforms all other methods.

5 An Efficient Image Super Resolution Model Using Generative Adversarial Networks

5.1 Introduction

The deep GAN models utilized in the context of SR possess a substantial depth, resulting in considerably high computational complexity. In this chapter, we present a novel and efficient GAN based model (SRINet) for image SR problem to overcome large network depth and high computational cost problems. We use complex filter structure to build the generator architecture. This structure is constituted using large and small convolution windows. The large windows assist the network to learn hierarchical features, whereas small windows help to extract local information from the image. In addition, dual stage enhancement approach is incorporated in the generator which leads the network to learn precise mapping from a LR feature space to generate a SR image. Residual learning based on dense skip connections is employed in generator to increase the feature learning capability of model. Our model exhibits superior performance compared to SOTA methods, yielding enhanced results.

Main contributions of the proposed model are threefolds:

1. In this chapter, we introduced a novel two stage progressive upscaling GAN, SRINet, for image SR. The generator component of our proposed model is partitioned into two distinct stages, wherein each stage performs a $\times 2$ enhancement, thereby enabling a final super-resolution factor of $\times 4$. Learning capability of our network increases by using progressive upscaling process. Training process is eased by introducing residual and dense skip connections between these two stages to eliminate the vanishing gradient problem.
2. A High Feature Generation Block (HFG-B) is introduced in the generator architecture consisting of cascaded modified inception blocks (MIBs). MIB assists the network to learn hierarchical features and primitive features present globally and locally in an image, respectively to learn more precise representations.

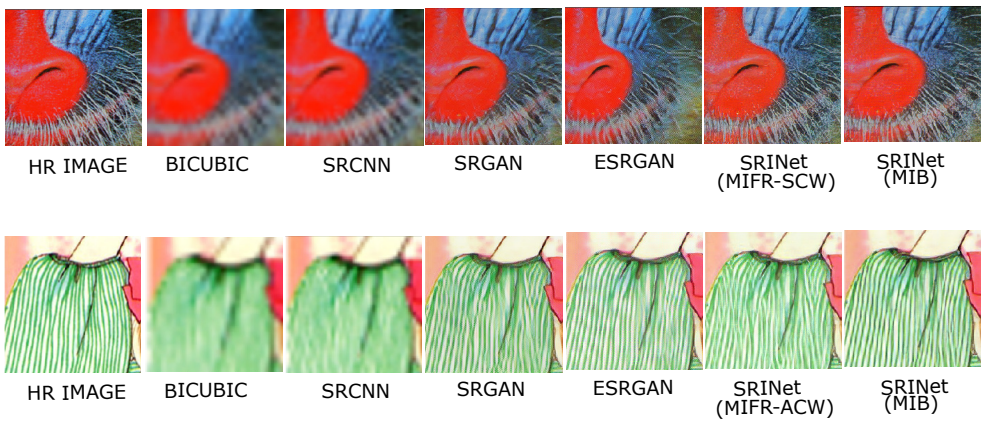


Figure 5.1: Visual results on the 'baboon' image from SET14 and 'YumeiroCooking' image from MANGA109 test dataset; upscaled by $\times 4$ factor.

3. Experiments on five benchmark datasets reflect better efficiency and superior performance over SOTA.

5.2 Methodology

5.2.1 Revisit Inception architecture

Inception model was introduced by Szegedy et al. [154] for image classification and detection. The fundamental inception block of the inception architecture is represented in Figure (5.2a). This architecture tries to overcome two major drawbacks in deep learning models:

- Over-fitting,
- Excessive usage of computational resources.

The key point in the inception architecture is to determine how the dense components present in convolutional neural networks can approximate the most favorable sparse structures. In inception architecture, complex filters are used instead of linear filters to foster the learning capability of model. These complex filters introduce non-linearity by following the structure of multi-layer perceptrons of 1×1 filter size; therefore, these structures perfectly fit in the CNN architecture. Global average pooling layer is used in the inception architecture abolishing the need of fully connected convolutional layer. This layer decreases the count of parameters used in the architecture, thus diminishing two major problems (over-fitting and high computational cost) of deep learning algorithms.

[155] further explored inception architecture to reduce computation cost and optimize memory usage. They presented an efficient way to scale up the convolution networks through optimization. By utilizing the parallel structure and dimension reduction, they redesigned the inception block through following modifications:

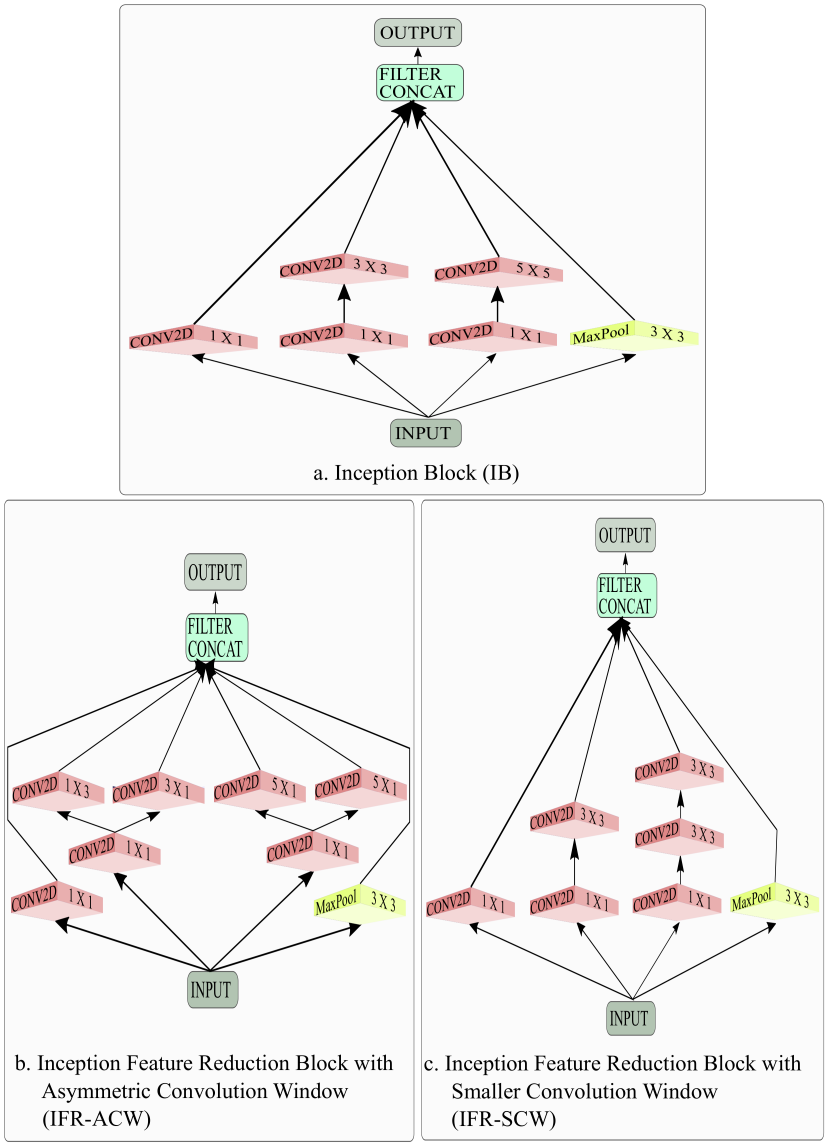


Figure 5.2: a. Fundamental Inception block (IB) b. Inception Feature Reduction Block with Asymmetric Convolution Window (IFR-ACW) c. Inception Feature Reduction Block with Smaller Convolution Window (IFR-SCW)

1. Asymmetric structure is suggested in place of symmetric structure. Suppose, we are using a convolution window of size 3×3, this window is replaced by stacking windows of 1×3 and 3×1 size (refer Figure (5.2b)). Thus reducing the number of computations from 9 to 6 for each convolution layer.

2. In place of using a single convolution layer with large convolution window, two layers are superimposed with smaller convolution window (refer Figure (5.2c)). This means that convolution window size of 5×5, i.e. 25 calculations is replaced by two superimposed layers with convolution window size 3×3.

5.2.2 Proposed model

In proposed model, SRINet, we use progressive upscaling approach for the generator architecture. SRINet uses two stage upscaling process (refer to Figure 5.3) to increase resolution of an image with an upscaling factor of 4 (each stage producing a $\times 2$ scaled output).

Each stage of SRINet is divided into three blocks: 1. Elementary Feature Extraction Block (EFE-B), 2. High Feature Generation Block (HFG-B), and 3. Reconstruction Block. To demonstrate the working of architecture, we denote convolution layer as Θ. Features obtained from first block consisting of a single convolution layer are given by F_0 eq. (5.1).

$$\mathcal{F}_0 = \hbar_{\Theta}(I_{LR}) \tag{5.1}$$

where convolution operation performed on the input image is represented by $\hbar_{\Theta}(.)$. This block extracts the primitive information from a low resolution image. Next, $\mathcal{F}_o$ is applied to the High Feature Generation Block (HFG-B), resulting in eq. (5.2).

$$\mathcal{F}_1 = \hbar_{hfg}(\mathcal{F}_0) \tag{5.2}$$

where $\mathcal{F}_1$ are the features obtained from the HFG-B. The operations performed within this block are represented by $\hbar_{hfg}(.)$. High level features obtained from the second block are passed to the reconstruction block comprises of a subpixel layer followed by the ReLU. Feature maps obtained from subpixel layer are applied to a convolution layer followed by a LeakyReLU activation function refer eq. (5.3)

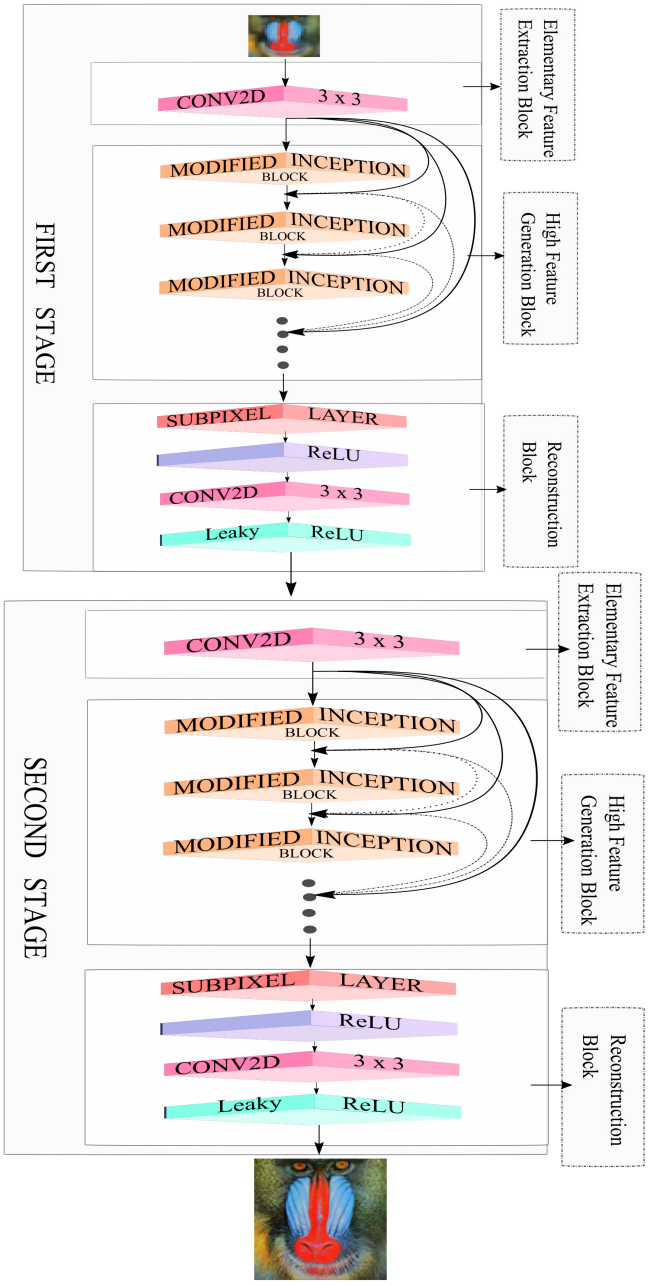


Figure 5.3: Generator architecture (SRINet)

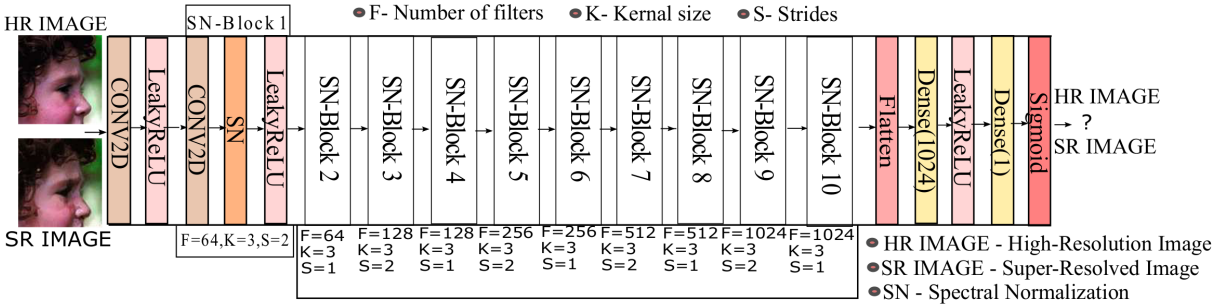


Figure 5.4: Discriminator architecture

$$\mathcal{F}_{2x} = \hbar_{rb}(\mathcal{F}_1) \tag{5.3}$$

$\mathcal{F}_{2x}$ are the output features obtained from the reconstruction block and $\hbar_{rb}(.)$ is the operation performed in this block. The upscaled output features generated from this block are applied further in second stage.

This stage has three blocks similar to the first stage with the difference in the size of the feature maps (double than previous stage). The final image obtained from the second stage is the SR $\times 4$ image (refer eq. (5.4)).

$$\mathcal{F}_{4x} = \hbar_{SB}(\mathcal{F}_{2x}) \tag{5.4}$$

where $\mathcal{F}_{4x}$ and $\hbar_{SB}(.)$ denote a final output image and operations performed in the second stage, respectively.

Two stage upsampling process provides superior perceptual results as compared to the models with single upscaling stage (refer section 4.3). This architecture mimics the frequency texture along with very fine details present in an image by increasing the learning capability of the model. Discriminator network (proposed by Ledig et al. [69]) is fed with the synthetic images from generator component (I_{gn}) and the HR images (I_{gt}) (refer Figure 5.4). This network is trained in such a way that it is capable of differentiating between these two images. The generator network is trained to produce the synthetic images similar to the HR images. So, the overall training process of these two networks will help to generate images with very HR.

5.2.3 High Feature Generation Block

In this section, we are going to illustrate the most essential part of SRINet; i.e., High Feature Generation Block (HPG-B). HPG-B consists of cascaded Modified Inception Blocks (MIBs). Output features obtained from the first MIB block are applied as an input to the second MIB and so on (up to the eighth block). Additionally, residual dense skip connections are introduced in the network to eliminate vanishing gradient problem. Output features obtained from the EFE-B are also merged with the output features of each MIB block to reduce the requirement of long term memory depedency. Also, the MIB blocks are densely connected with each other as shown in Figure 5.3. These connections provide fast learning competency to the network with high frequency feature maps.

5.2.3.1 Modified Inception Block

Modified Inception Block (shown in Figure 5.5) is based on the fundamental inception block of inception architecture. The salient parts of an image have huge inequality in terms of size in different images. Inception architecture overcomes this problem by using different convolution size windows on the same level. Convolution layers with large window size are used to extract the hierarchical features of an image. And convolution layers with small window size are used to extract the locally distributed features of an image. Additionally, convolution layer with 1×1 window size is introduced before the large filter size convolution layers to reduce the computational complexity. This limits the input channels for the model and hence reduce the number of parameters for the training. Maxpooling layer is also present in the fundamental inception architecture (refer to figure 5.2). Pooling layer is used in classification problems to suppress the irrelevant features and identify an object in an image. But in image super resolution, we are generating the unavailable data by using the available pixels. Therefore, pooling layer has no significance is this problem. Pooling layer produces artifacts in the SR system (discussed in detail in section 5.3.3.2 and Figure 5.6). Also, the number of parameters and computational complexity is increased to enormous amount by using maxpooling layer in the inception block. Therfore, we modified the fundamental inception architecture (refer figure 5.5a) to make it suitable for image super resolution and named it as Modified Inception Block (MIB). In place of ReLU activation function we use LeakyReLU activation function. LeakyReLU activation helps the network to achieve good local minimum as compared to the ReLU activation function. For negative inputs, ReLU gives zero value which in turns makes the gradient value zero. So, in some cases the network is not able to achieve good local minimum while using ReLU activation function.

The MIB, shown in Figure 5.5a, has three parallel pipelines of convolution layer depicted by PL_1, PL_2 and PL_3. The feature maps ($\mathcal{F}_0$) obtained from eq (7.1) are passed through these

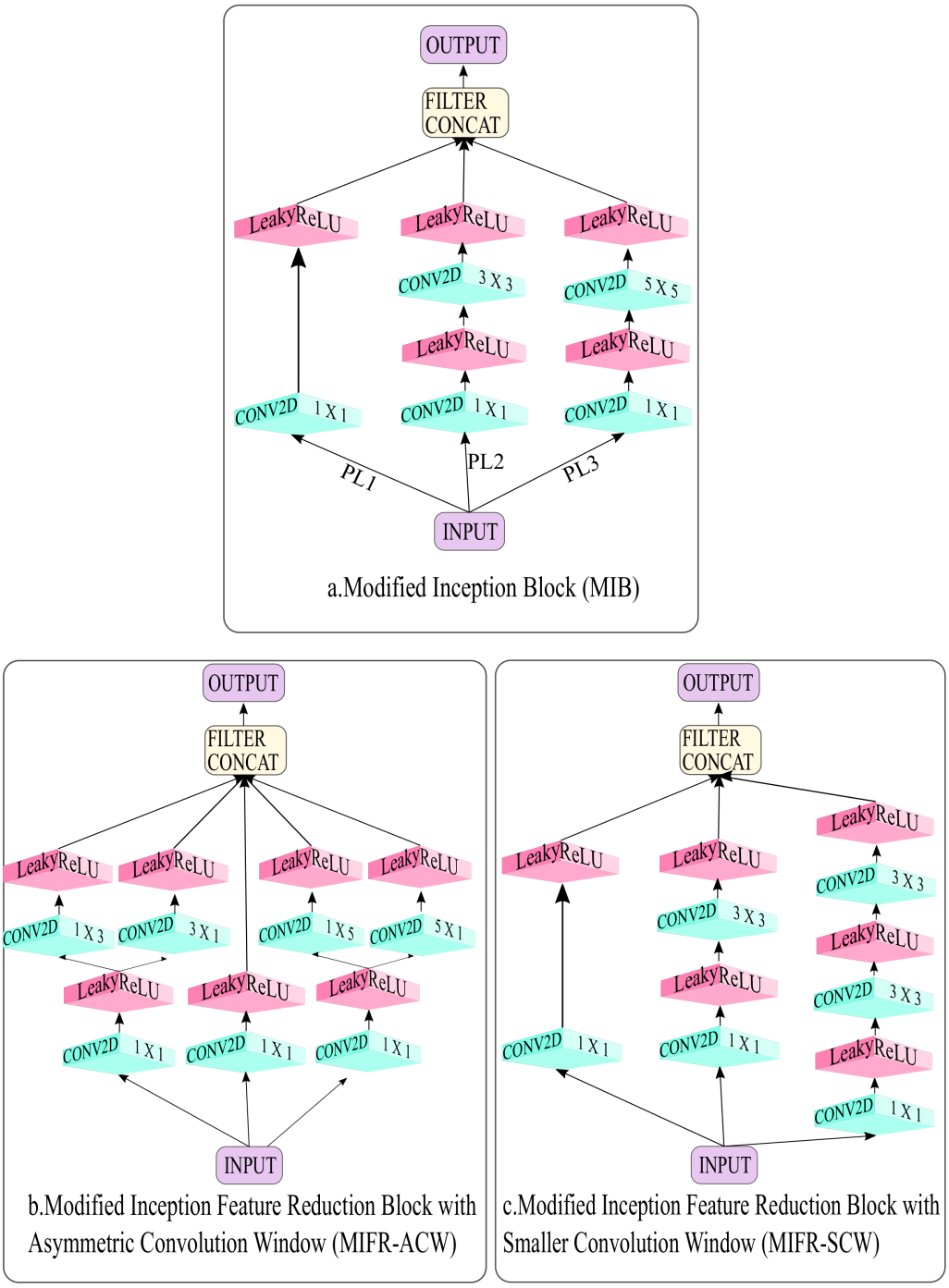


Figure 5.5: a. Modified Inception Block (MIB) b. Modified Inception Feature Reduction Module with Asymmetric Convolution Window (MIFR-ACW) c. Modified Inception Feature Reduction Module with Smaller Convolution Window (MIFR-SCW)

pipelines to produce features x_1, x_2, x_3, x_4 and x_5 (refer eqs. (5.5), (5.6), (5.7), (5.8) and (5.9).

$$x_1 = \Phi[\mathcal{F}_0\Theta_{(1\times1)}W_1 + B_1] \mid_{PL_1} \tag{5.5}$$

$$x_2 = \Phi[\mathcal{F}_0\Theta_{(1\times1)}W_2 + B_2] \mid_{PL_2} \tag{5.6}$$

$$x_3 = \Phi[x_2\Theta_{(3\times3)}W_3 + B_3] \mid_{PL_2} \tag{5.7}$$

$$x_4 = \Phi[\mathcal{F}_0\Theta_{(1\times1)}W_4 + B_4] \mid_{PL_3} \tag{5.8}$$

$$x_5 = \Phi[x_4\Theta_{(5\times5)}W_5 + B_5] \mid_{PL_3} \tag{5.9}$$

where LeakyReLU layer and convolution layer is represented by Φ and $\Theta_{w\times w}$, respectively. Subscript w is showing the kernel size of each convolution layer. $\mid_{PL_m}$ indicates the processing in mth pipeline ($m \in [1, 2, 3]$).

The feature maps x_1, x_3 and x_5 obtained from eqs. (5.5), (5.7) and (5.9) are concatenated to produce the output feature maps x (refer eq. (5.10))

$$x = concat[x_1, x_3, x_5] \tag{5.10}$$

These features maps are applied as an input to the next MIB block and so on.

To reduce the memory usage and computational cost, we modify the inception v3 architectures (refer b and c part of Figure (5.2)) for super resolution problem as shown in Figure 5.5 (b,c).

We use three different approaches for HFG-B in the generator to implement SRINet:

- Modified Inception Block (MIB)

- Modified Inception Feature Reduction block with Asymmetric Convolution Window (MIFR-ACW)
- Modified Inception Feature Reduction block with Smaller Convolution Window (MIFR-SCW).

The other two blocks (EFE-B and reconstruction block) remain same for these three generators.

5.2.4 Loss Function

The final output image is obtained by optimizing the loss function (L^{SR}) over K number of training samples. Loss function (L^{SR}) is formed based on Ledig et al. [69] by adding two losses: feature-based contextual loss and adversarial loss.

$$L^{SR} = L^{SR}_{vgg} + L^{SR}_{adv} \tag{5.11}$$

Pre-trained VGG-19 model is used for extracting the feature maps for the ground-truth images and the output images. Convolution layer, before the last max pooling layer of Pre-trained VGG-19 model, is used to extract the features maps of ground-truth images and output images. The Euclidean distance is calculated between the feature maps of these two images. Feature-based contextual loss is given by Eq (5.12).

$$L^{SR}_{vgg/i.j} = \frac{1}{W_{i,j}H_{i,j}} \sum_{a=1}^{W_{i,j}} \sum_{b=1}^{H_{i,j}} (\psi_{i,j}(i^{gt})_{a,b} - \psi_{i,j}(G_{\Theta_g}(i^{lr}))_{a,b})^2 \tag{5.12}$$

Here, W and H are the dimensions of the extracted feature maps. $\psi_{i,j}(i^{gt})$ and $\psi_{i,j}(G_{\Theta_g}(i^{lr})$ represents the feature maps extracted from VGG-19 model for ground-truth images and the output images produced from the generator respectively.

To calculate adversarial loss (L^{SR}_{adv}, refer eq. 5.13, generator component is added to loss function. Addition of this component guides the network to generate images similar to the real images by misleading the discriminator. Mathematical formulation of adversarial loss is shown in Eq. (5.13).

$$L^{SR}_{adv} = \sum_{k=1}^{K} -logD_{\Theta_d}(G_{\Theta_g}(i^{lr})) \tag{5.13}$$

Here, Θ_d and Θ_g are the optimization parameters (weights and biases) for discriminator and generator network, respectively. $D_{\Theta_d}(G_{\Theta_g}(i^{lr})$ is the probability that the generated image is the real image.

5.3 Experiments

5.3.1 Datasets

From the imagenet dataset, 80,000 images are selected randomly to train the dataset. Experiments are performed on five benchmark SR datasets: Set5 [148], Set14 [150], BSD100 [156], Urban100 [157] and Manga109 [158] to validate the performance of the proposed architecture.

5.3.2 Training settings and implementation details

The model is evaluated for the upscaling factor of $\times 4$ between the ground-truth and low resolution image. Crop size for the ground-truth image is 192×192. Gaussian noise is added to the ground-truth image and then bicubic kernel is used to downsample it by a factor of $\times 4$ to generate a low resolution image of size 48×48.

3×3 window size with 64 channels are used for EFE-B convolution layer. In the HFG-B, number of channels used in each convolution layer are fixed to 64. Each HFG-B consists of 8 MIBs. The subpixel layer in the reconstruction block has 64 channels with 5×5 filter size. The convolution layer of the first stage reconstruction block has 64 channels with 3×3 window size. The convolution layer of the second stage reconstruction block has 3 channels with 3×3 window size to get a RGB image at the output. Thus, the total number of convolution layers used in the proposed generator are 86. LeakyReLU with $\alpha = 0.2$ is used as an activation function in both generator and discriminator. β_1, β_2 has values 0.9 and 0.999, respectively with learning rate = 0.0001 and batch size of 16; the network is optimized with Adam optimizer. The batch size is 16. From training dataset, samples are produced randomly to carry out training in batches. Total number of steps used for training the model are 10^5 and number of steps per epoch are 5000. Loss functions, to minimize the difference between HR and LR image, is L^{SR} (refer section 3.3). Discriminator and generator networks are alternatively updated until the model is fully trained. Each of these networks are held constant while training the other. Two most common quality analysis metrics PSNR (Peak Signal to Noise Ratio) and SSIM (Structural Similarity Index Measurement) are used to evaluate the proposed model. Both these metrics are calculated on the Y-channel for fair comparison with the former works. To measure the perceptual quality of images, we used Visual Information Fidelity (VIF) metic [159] which rely on natural scene statistics and human visual system.

Table 5.1: Contribution of different components: Performance investigation on SET5 due to different components present in our architecture.

	a	b	c	d	e	f
Single stage	✓	✓	✓			
Two stagges				✓	✓	✓
MaxPooling layer	✓	✓		✓	✓	
ReLU	✓			✓		
LeakyReLU		✓	✓		✓	✓
PSNR	12.83	14.24	22.78	24.23	24.56	27.64

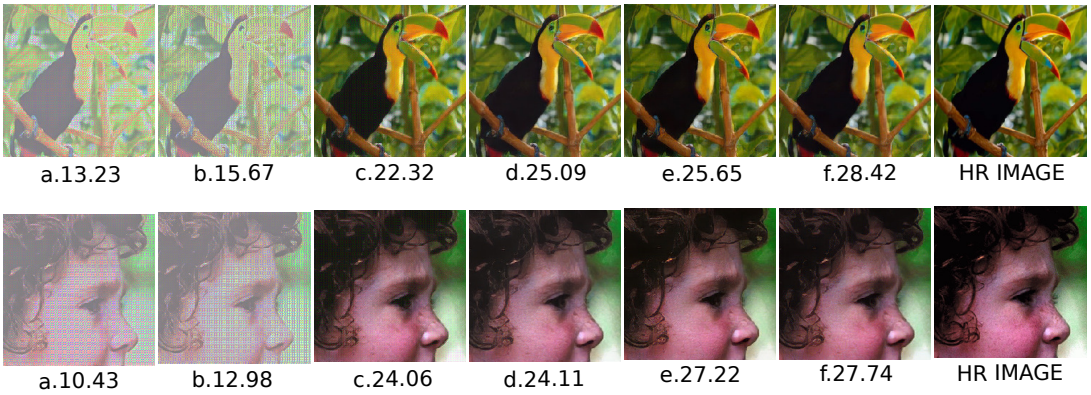


Figure 5.6: Performance investigation on SET5 due to different components present in our architecture.

5.3.3 Ablation study

5.3.3.1 Analysis of reconstruction module

We studied the effect of reconstruction block at two different positions. Firstly, we used reconstruction block at the end of the architecture as done in many earlier researches, such as, SRGAN [69] and ESRGAN [84] etc. Then, we applied the progressive upscaling approach to our model, as shown in Figure 5.3. Results obtained (refer Figure 5.6) after applying these two approaches supports our claim that progressive upscaling is better than the single stage upsampling process. As this approach provides high learning capability to our model and the model is able to mimic the fine texture details of an image.

5.3.3.2 Analysis of maxpooling layer in inception module

We tried to solve the SR problem by using fundamental inception block (refer Figure 5.2a) in the HFG-B. Number of parameters for this architecture are very high which further increases the time per epoch for training. With the huge computational cost and high memory usage, this model was not able to perform well for the SR problem as shown in Figure 5.6. The maxpooling layer produces artifacts in the image. This layer is used to suppress the irrelevant

features while performing tasks like image classification and detection. In super resolution data is generated with the help of available pixels, thus this layer has no significance in SR task. Therefore, we remove maxpooling layer from the inception module and examine the results. After removing maxpooling layer from the IB, number of parameters and computational cost for the architecture is reduced to enormous rate as with improvement in the perceptual quality of images. So, qualitative and quantitative results obtained with and without the max pool layer support out claim that our network is performing better after removing the max pool layer from the inception block for image super resolution task.

5.3.3.3 Analysis of activation function in inception module

We studied the effect of activation function in the inception block. In basic inception block ReLU activation function is used after the convolution layers to add non-linearity. Sometimes, ReLU activation function is not able to achieve local minimum because it generates zero value for the negative inputs which turns the gradient value to zero. Therefore, we removed the ReLU activation function and use LeakyReLU activation after the convolution layer in the proposed generator. Comparison results are shown in Figure 5.6 with ReLU activation function and LeakyReLU activation function. From the experiment results, it is clear that after employing LeakyReLU activation function in the inception block there is significant increase in the performance of the model as compare to the model with ReLU as activation function.

5.3.4 Comparison with state-of-the-art methods

We compared our model with bicubic interpolation and other SOTA methods: SRCNN [67], ESPCNN [68], SRGAN [69], ESRGAN [84], GSR-DDNet [85] and RNAN [93]. We presented the qualitative and quantitative analysis on benchmark datasets. For the quantitative analysis, average PSNR and average SSIM is calculated on benchmark datasets presented in table 5.2. For fair comparison, the PSNR and SSIM values are calculated on the Y-channel similar to the former works. For the qualitative analysis perceptual results are shown in Figures 5.7, 5.8, 5.9, 5.10, 5.11, 5.12 and 5.13 with their VIF scores. As shown in table 5.2, our model outperforms over all the GAN based SR models in terms of PSNR ans SSIM for all the datasets except BSD100.

Our model has achieved second highest average PSNR and highest average SSIM for Set5. Perceptual results evaluated on Set5 are shown in Figure 5.7. From the figure, it is clear that CNN based methods produced very smooth result. The competing GAN based methods are also unable to recover the contours present in the image. Only the proposed method produces results similar to the ground-truth image with the highest VIF score.

Table 5.2: Performance comparison on the basis of average PSNR (dB) and average SSIM for various SR methods on various test datasets (Set5, Set14, BSD100, Urban100 and Manga109) (First and second highest are Bold) with scale $\times 4$ factor.

Dataset	Set5		Set14		BSD100		Urban100		Manga109	
Metric	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	23.41	0.757	24.42	0.724	23.04	0.618	20.01	0.565	20.32	0.692
SRCNN [67]	24.32	0.793	24.80	0.771	23.24	0.615	20.39	0.588	20.49	0.703
ESPCNN [68]	26.55	0.811	25.31	0.793	24.93	**0.629**	21.43	**0.691**	21.38	0.724
SRGAN [69]	26.71	0.812	26.05	0.827	24.08	0.510	22.54	0.672	23.04	0.791
ESRGAN [84]	25.80	0.834	25.73	0.813	23.60	0.518	21.88	0.650	22.21	0.755
GSR-DDNet [85]	26.72	0.820	25.86	0.822	23.92	0.523	21.99	0.643	22.13	0.753
RNAN [93]	**29.35**	**0.868**	**30.02**	**0.851**	**26.31**	**0.691**	**24.74**	**0.703**	**26.01**	**0.814**
TwoFold SRINet (MIFR-ACW)	26.73	0.805	27.84	0.699	**24.46**	0.543	21.15	0.591	21.99	0.732
TwoFold SRINet (MIFR-SCW)	26.87	0.799	28.30	0.733	23.54	0.539	21.93	0.580	21.53	0.701
TwoFold SRINet	**27.64**	**0.870**	**28.90**	**0.840**	24.42	0.608	**22.71**	0.675	**23.29**	**0.792**

For Set14, our model has second highest PSNR and SSIM. Comparison results for the qualitative analysis are presented in Figures 5.8 and 5.9 with their PSNR, SSIM and VIF scores for Set14. CNN based methods are generating blurry effect in the output images. In Figure 5.9, only our method is able to recover the edges and the lines present in the upper part of image (crown) with higest VIF score.

For BSD100 dataset, the proposed model is third highest for average PSNR and SSIM. But the perceptual results on our method are better than all the competing methods (refer Figures 5.10 and 5.11). In figure 5.10, most of the methods are producing blurry effect on the background and unable to recover the sharpness of an image. Whereas visibility of lines in proposed method is very clear with sharp details which is reflected in the VIF scores also.

Our model has achieved second highest PSNR and third highest SSIM for Urban100 dataset. In Figure 5.12, small textural details and the horizontal lines are blemished in almost all the competing methods. The proposed method recovers small textured details present on the bricks with clear horizontal lines and maximum VIF score, which shows our model's superiority over other competing methods.

For Manga109 dataset, our model has achieved second highest PSNR and SSIM. Perceptual results for this dataset are shown in Figure 5.13. From the results, it is clear that no competing methods are able to recover the exact font of letters and background details are also missing in the output images. Our model is producing results similar to the output image with clear

background details and sharp font of letters. From the analysis, it is clear that perceptually our model is performing better than other models.

Figure 5.7: Perceptual results with their PSNR/SSIM/VIF score on the 'butterfly' image from the SET5; upscaled by $\times 4$ factor.

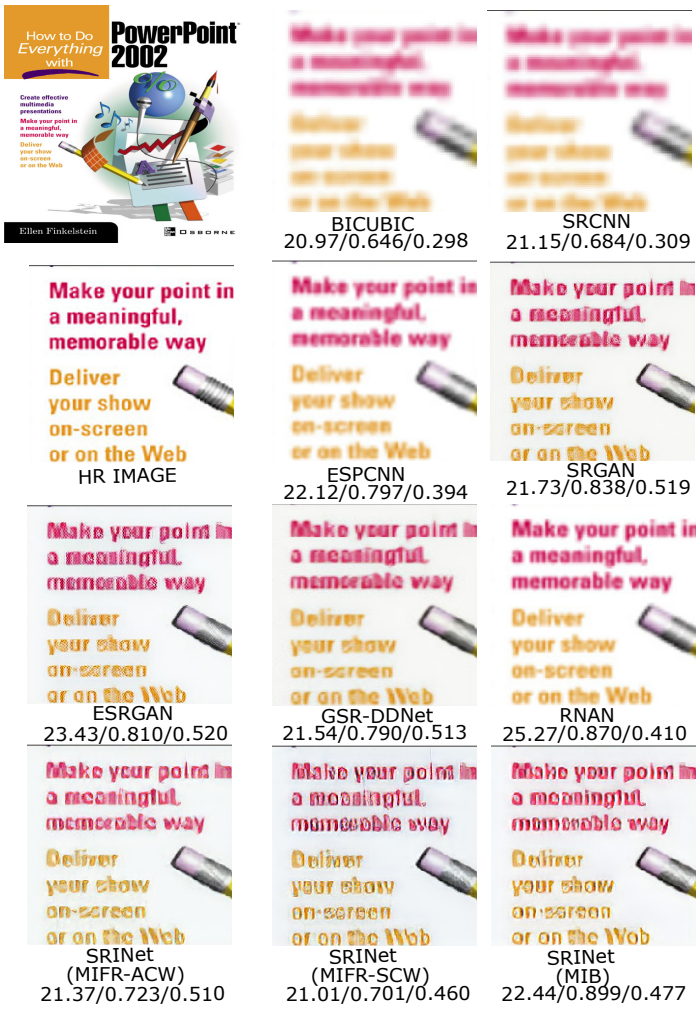


Figure 5.8: Perceptual results with their PSNR/SSIM/VIF score on the 'ppt3' image from the SET14; upscaled by $\times 4$ factor.

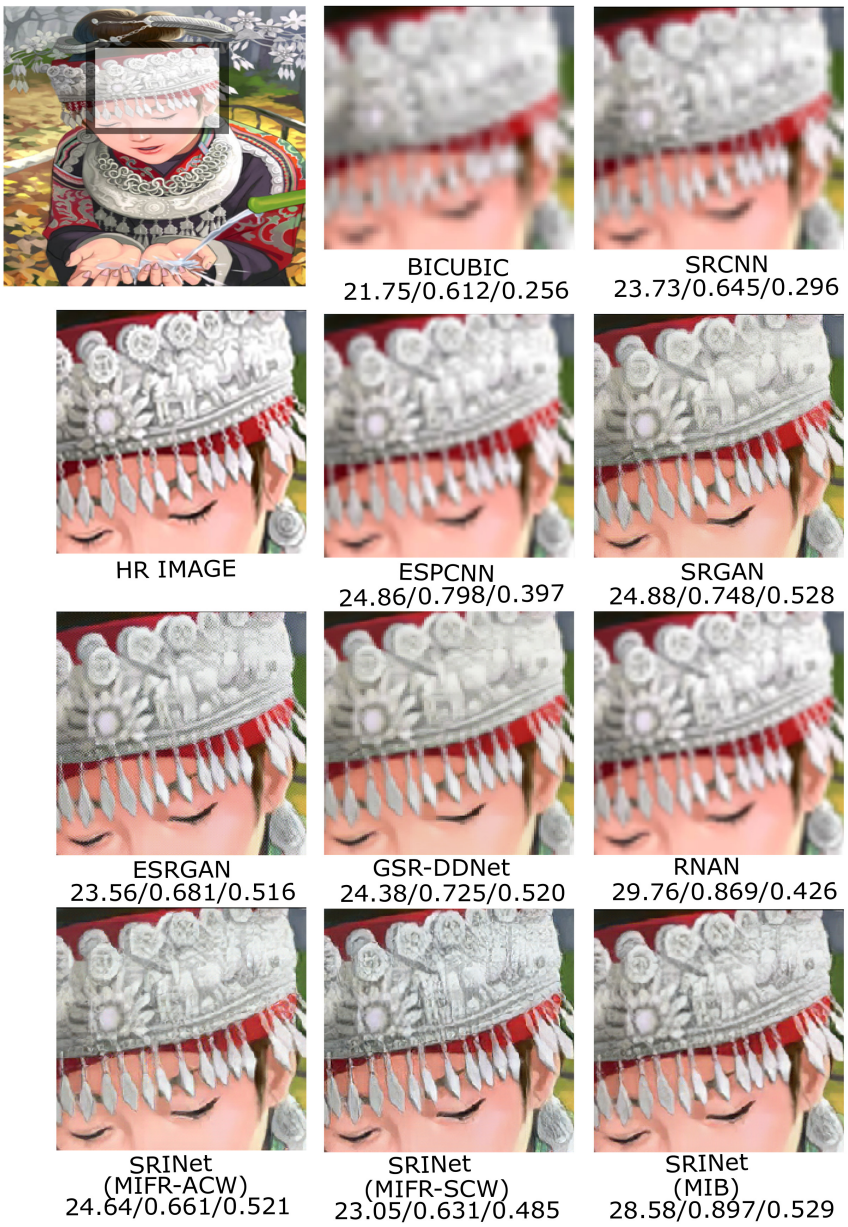


Figure 5.9: Perceptual results with their PSNR/SSIM/VIF score on the 'comic' image from the SET14 test dataset; upscaled by $\times 4$ factor.

Figure 5.10: Perceptual results with their PSNR/SSIM/VIF score on the '101085' image from the BSD100; upscaled by $\times 4$ factor.

Figure 5.11: Perceptual results with their PSNR/SSIM/VIF score on the '102061' image from the BSD100 test dataset; upscaled by $\times 4$ factor.

Figure 5.12: Perceptual results with their PSNR/SSIM/VIF score on the '076' image from the URBAN100 test dataset; upscaled by $\times 4$ factor.

Figure 5.13: Perceptual results with their PSNR/SSIM/VIF score on the 'YumeNoKayoiji' image from the MANGA109 test dataset; upscaled by $\times 4$ factor.

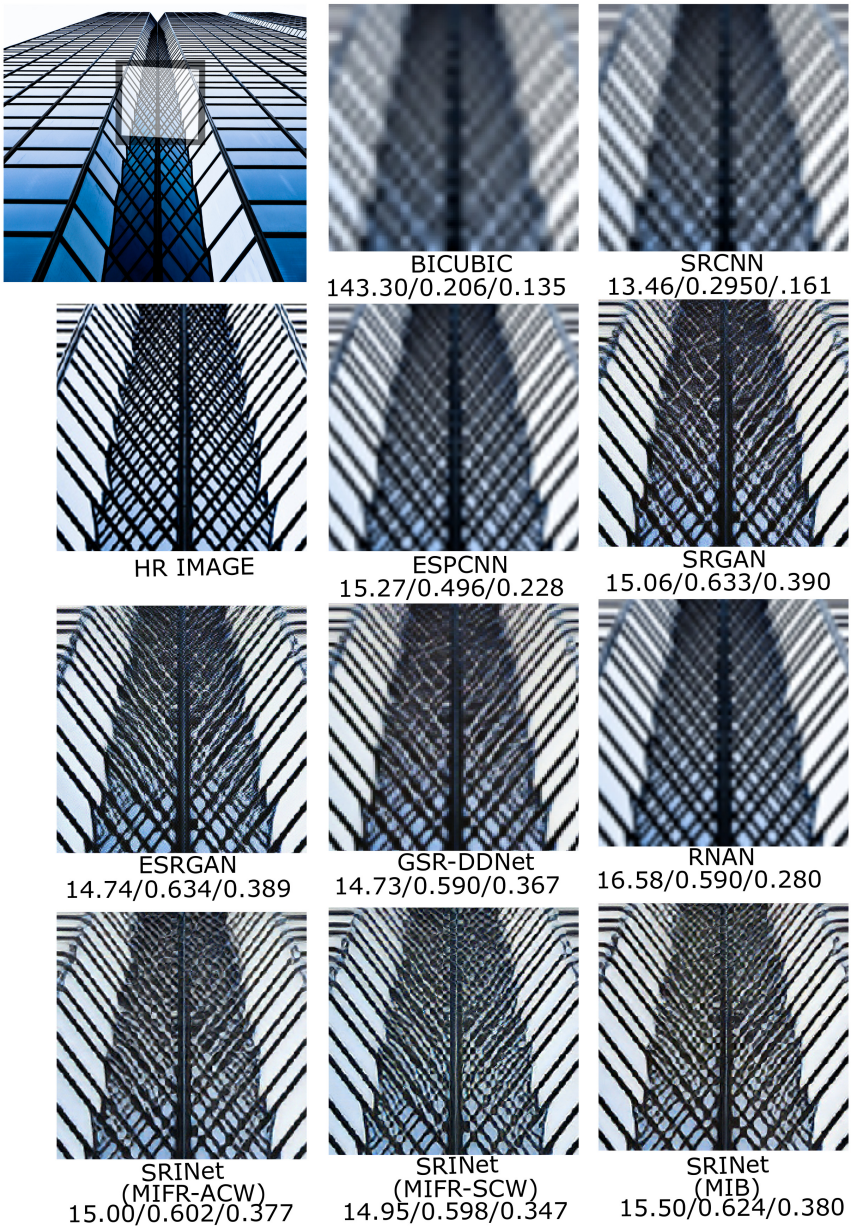


Figure 5.14: Failure case on the '067' image from the URBAN100 test dataset for image super resolution; upscaled by $\times 4$ factor.

5.3.5 Limitations

Our model shows good performance for SR task by producing very good perceptual and qualitative results as compare to other state-of-the-art methods. In few cases quantitative results produced by our model hace lower values than the CNN based SR methods (refer table 5.2). In few images of Urban100 dataset, our model, like all other competing models, struggles to visualize subtle details (refer Figure 5.14).

5.4 Conclusion

Current chapter presents a novel GAN Based SRINet model. This model alleviates the use of linear filters and integrates the complex filter structures in the network to approximate most favorable sparse structures. Dense skip connections are introduced in the architecture. This approach increases the learning capability of network while retaining its computational complexity. In addition, progressive upscaling approach is used, where image enhancement is done in two stages, to sustain high frequency components, allowing us to produce output images with fine texture details. The model is able to produce better perceptual results than the other SOTA methods.

6 Frequency Aware and Semantic Structural Constraint Based Face Hallucination System

6.1 Introduction

Particularly, applying SR on faces is known as Face Hallucination (FH). High resolution facial images are widely required in many computer vision applications, such as facial emotion detection [160], pedestrian re-identification [161], facial alignment [162], face recognition [163] and face identification [164].

Based on deep learning, researchers have effectively applied numerous algorithms to solve face SR problem [165]. However, the first drawback of majority of these algorithms is that they rely on two dimensional facial priors to recover structural details of facial images, such as, Grm et al. [166] proposed a face SR algorithm, where the identity priors are incorporated at multiple stages of CNN to perform image hallucination. Progressive facial attention loss is proposed by Kim et al.[167] to incorporate facial attributes in the generated SR face images. Yu et al.[168] used coarse SR image from the intermediate stage to extract heatmaps by passing that image in an UNet architecture. Extracted heatmaps are then merged with coarse SR image to generate final SR image. All the above mentioned methods used 2D priors to guide the deep learning models to generate the HR face images. The information extracted from the 2D priors is only capable of incorporating global features in an output image. Still the local features or the subtle structural details like skin irregularities, wrinkles and depth details are missing in the final output image. To embed the above mentioned features in the generated image, we have proposed a progressive Face Hallucination (FH) network. An auxiliary subnetwork is employed in the proposed FH network by using 3D Morphable Models (3DMMs) [169] to embed structural information in the output image. 3DMMs are the three dimensional meshes of face images used to reconstruct a 3D image from its 2D counterpart using shape and texture models of Principal Component Analysis (PCA). To add the semantic structural constraint to proposed FH model, we are utilizing the PCA shape model of 3DMM [170]. The shape components of PCA shape model constitute different face parameters, like, the first shape

component signifies the shape of face (slim, chubby or round etc). And the secondary shape components accounts for the more finer details (wrinkles, face irregularities, face depth etc) of face images. These three dimensional meshes with face parameters are rendered as two dimensional points on the face image. The obtained 3D points fitted on 2D images act as a target images for our auxiliary network. This auxiliary network act as a supervision network to add structural constraint to our face hallucination network.

Second drawback of Super resolution methods based on generative adversarial networks [69, 84] is that the quantitative results produced by these methods have less values as compared to other deep learning based methods due to missing frequency details. To overcome this drawback, a sub-network comprised of an auto encoder is added along with the proposed FH network to explicitly add high frequency facial details from an face image. Discrete Cosine Transform (DCT) based feature maps with high resolution are generated using this network. Frequency domain loss is calculated between the generated DCT feature maps and ground truth DCT feature maps which is used to train the sub-network. This loss function will guide our FH network to produce output images with high quantitative values. To embed high feature DCT maps in the FH network, we used an Inverse Discrete Cosine Transform (IDCT) block. This IDCT block converts the frequency domain feature maps into the spatial domain feature maps. Then the converted feature maps are merged with the output feature maps of FH network, supervising our network to produce images with high frequency details.

Main contributions of presented work are as follows:

1. GAN based progressive face hallucination network is proposed. The generator network in the proposed network comprises of FH network and two sub-networks, assisting FH network to generate high resolution face image. The FH network consists of Hierarchical Feature Extraction Module and Computationally Efficient Channel based Attention Module. The hierarchical feature extraction module helps the model to learn hierarchical as well as primitive information present in the image. Where as channel based attention block is used to add channel-wise attention in the network and reduce the computational complexity of the network.
2. First sub-network produces high resolution DCT feature maps. This network supervises FH network to produce images with high frequency details. Proposed frequency domain based loss, assists our face hallucination network to reflect the quality of resultant images. IDCT block is used at the end of this network to convert the frequency domain feature maps to spatial domain and merge them in the FH network.
3. An auxiliary sub-network generates a high resolution 2D image with 3D parameters fitted on it. This network adds, structural constraint to our FH network, producing face images with semantic facial details, like skin irregularities, wrinkles and depth information etc.
4. Experiments on facial benchmark datasets reflect superior performance over recent SOTA.

6.2 Methodology

As depicted in the Figure 6.1, generator network of the proposed architecture consist of three branches:

i) Progressive face hallucination branch (PFH-B), powerfully built with a combination of cascaded hierarchical feature extraction module and channel based attention module,

ii) Semantic structural constraint branch (SSC-B), serves as a supervision network by constraining PFH-B to generate resultant images with three dimensional parametric feature information, and

iii) DCT based auto encoder branch (DCTAE-B), compelling PFH-B to produce images with high frequency details.

Basically, the objective of proposed face hallucination model is to find the mapping function $\mathcal{F}_{\theta_{fh}}$ (refer eq. 6.1) to obtain a HR face image (HR_{fi}) from its LR counterpart (LR_{fi}).

$$\mathcal{F}_{\theta_{fh}} = LR_{fi} \rightarrow HR_{fi} \tag{6.1}$$

where, θ_{fh} are the parameters learned throughout the mapping process. To minimize the distance between HR_{fi} and its counter LR_{fi}, proposed face hallucination network employs the combination of pixel based loss, feature based loss, structured parametric loss and DCT based loss and updates the learnable parameters during the training process. The output face images generated by face hallucination network is fed to discriminator network [69] along with the HR face images. In the proposed architecture, discriminator is acting as a binary classifier, and trained in such a way that it classifies the ground-truth face images as label 1 and generated face images as label 0. On the contrary, generator is trained to trick discriminator by generating the face images similar to HR face images. Comprehensive training of both these networks will lead to the generation of HR face images.

Detailed explanation of components employed in the generator network are discussed next.

6.2.1 Progressive face hallucination branch

As illustrated in Figure 6.1, PFH-B uses a progressive upscaling technique where at every stage the input image is upscaled by the factor of 2. Further, each stage is divided into three phases. First phase is elementary characteristic extraction phase (ECE-P), where LR face image is passed through a convolution layers to extract the elementary characteristics of an image. The feature maps obtained from the first phase are applied to the second phase which is hierarchical

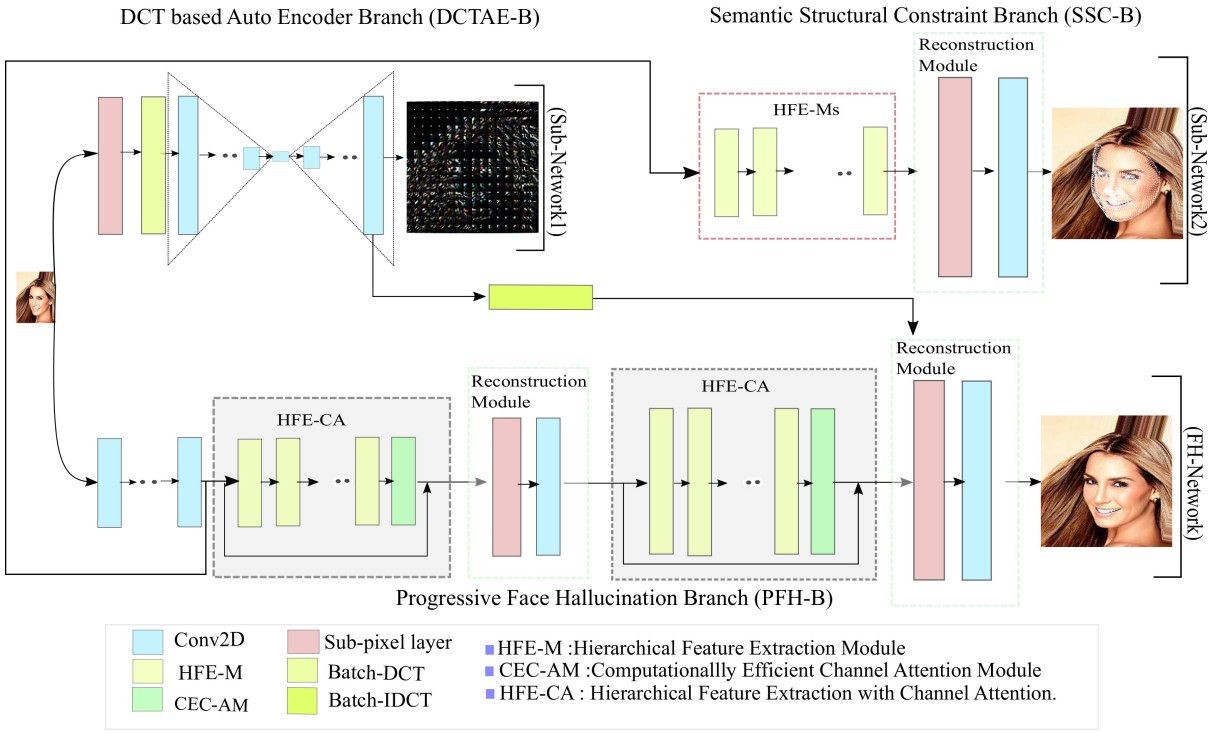


Figure 6.1: Proposed Generator architecture with three branches: 1) FH network- progressive face hallucination branch from where resultant output image is generated, 2) sub-network1- a DCT based encoder network to add high frequency components in the output image, and 3) sub-network2- semantic structural constraint branch to add 3D parametric information in the generated image.

feature extraction with channel attention mechanism (HFE-CA). This is the most crucial component of PFH-B. HFE-CA is composed of two main blocks - hierarchical feature extraction block and computationally efficient channel based attention module, which are explained in detail in the following subsection. Output feature maps obtained from the HFE-CA are applied to the reconstruction phase, where the features maps are upsampled by the factor $\times 2$ using subpixel convolution layer [68] to get the final output image.

6.2.1.1 Hierarchical feature extraction module

Hierarchical feature extraction module (HFE-M) is shown in Figure 6.2. Hierarchical feature extraction block (HFE-B) is the building block of HFE-M. To sustain the long term memory dependency, residual connections are used between the HFE-Ms. Total five HFE-Bs are used in each HFE-CA phase, where the output feature maps of first HFE-B are applied to the second block and so on.

The motivation of hierarchical feature extraction module (HFE-M) is taken from the inception architecture [154]. As depicted in the Figure 6.2, the same input is applied across three different convolution layers with different kernel sizes. Notion for utilizing this complex filter structure

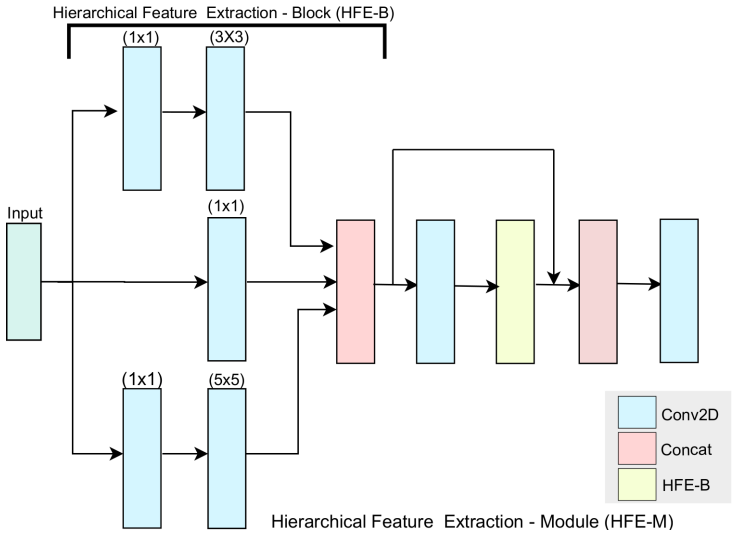


Figure 6.2: Hierarchical feature extraction module

across the input is to acquire local as well as global features of an input face image. Salient attributes like nose, eyes, lips, ears and wrinkles of face images have distinct sizes across an image. Therefore, to extract local attributes from an image, HFE-M uses convolution layers with small kernel sizes (1×1, 3×3). While the hierarchical and the global features are extracted using larger kernel sizes (5×5). Before the convolution layer with large kernel sizes (3×3, 5×5), convolution layer with 1×1 kernel size is used to curb the input channels and hence reducing the computational parameters of the architecture. All the convolution layers are followed by LeakyReLU activation function in order to introduce non-linearity in the model. The final feature maps are obtained by concatenating the individual feature maps obtained from each convolution layer with different kernel size.

6.2.1.2 Computationally efficient channel based attention module

As depicted in Figure 6.3, the motivation for computationally efficient channel based attention module (CEC-AM) is taken from mobilenet V3 model [171]. Rather than using regular convolution layer, depthwise separable convolution layers (combination of depthwise convolution layer and pointwise convolution layer) are used in this block. In depthwise convolution, for each channel in the feature space a single filter is applied and followed by 1×1 convolution using pointwise convolution layer to amalgamate the feature maps of depthwise convolution layer. This layer is preferred over the regular convolution layer due to its ability to use lesser number of computational parameters without affecting the functionality of traditional convolution layer. Second fundamental component used in this module is squeeze and excitation block [172]. The essence for using this block is to explicitly model the mutuality present between the channels of convolution feature maps, guiding the network to enhance the representational quality of output features. Thus, CEC-AM is used to incline the architecture's ability to assign the accessible

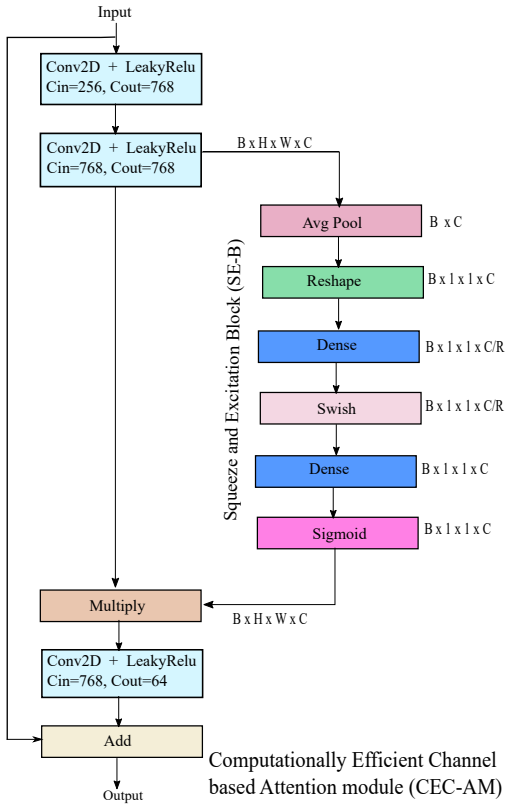


Figure 6.3: Computationally efficient channel based attention module

processing resources to the most essential information present in the input feature maps.

6.2.2 DCT based auto encoder branch (DCTAE-B)

As shown in Figure 6.1 (sub-network1), a discrete cosine transform based auto encoder is employed in parallel with PFH-B to incorporate frequency details in our SR network. Basically, following two contributions are proposed in this sub-network: 1) an auto encoder is employed in parallel with PFH-B to generate high resolution DCT coefficients from low resolution DCT coefficients. 2) DCT and IDCT blocks are defined in the network to transform data from spatial domain to frequency domain and vice versa.

The autoencoder takes the DCT coefficients of LR image as input and upsample it to the DCT coefficients of HR scale. The use of skip-connections ensure long passage of information in the network. We do not use any form of normalization in the network as it tends to produce artifacts in the image. The DCT to IDCT block helps in transforming the DCT coefficients back to spatial domain, and thus provides a common link between frequency and spatial domains. The output of AE is merged with the output of PFH-B. This serves the purpose of using DCT coefficients which have high frequency explicitly embedded in it.

6.2.2.1 DCT and IDCT module

To operate in the frequency domain, firstly the face images are converted from the spatial domain to frequency domain using DCT. For a single block, DCT is calculated by the formula given in equation 6.2 and DCT of ground-truth image is shown in figure 6.4

$$D_{a,b} = \frac{1}{\sqrt{2M}} \beta(a)\,\beta(b) \sum_{x=0}^{M}\sum_{y=0}^{M} i_{x,y} \cos\left[\frac{(2x+1)\,a\pi}{2M}\right] \cos\left[\frac{(2y+1)\,b\pi}{2M}\right] \tag{6.2}$$

here, block size is represented by M, image is denoted by i and pixel coordinates as x and y. a and b represents indexes of spatial frequency. Scale factor β (refer eq. 6.3) is used for transform to be orthogonal.

$$\beta(v) = \begin{cases} 1/\sqrt{2}, & \text{if } v = 1. \\ 0, & \text{otherwise.} \end{cases} \tag{6.3}$$

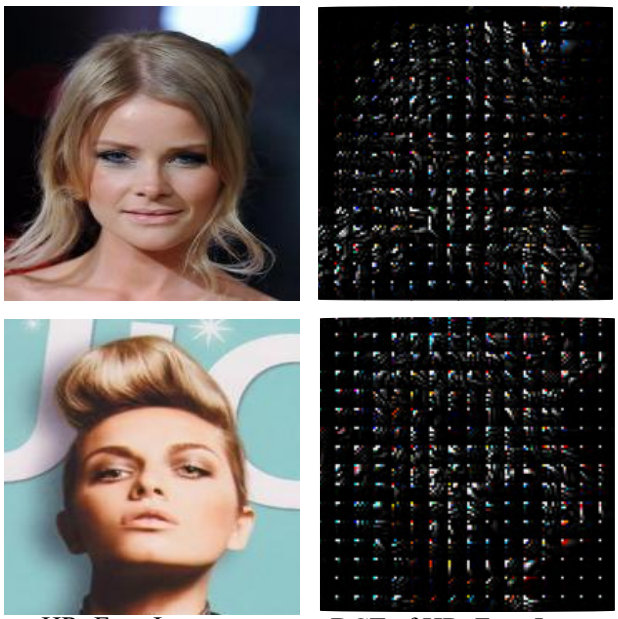
In order to establish a connection between the DCT based encoder network and main SR network, IDCT is used. IDCT transforms the frequency domain coefficients back to the spatial domain. The transformed values are then fused with the main SR network to get output images with high frequency details. The formula to calculate IDCT for single block is given by eq. 6.4

$$i_{x,y} = \frac{1}{\sqrt{2M}} \sum_{x=0}^{M}\sum_{y=0}^{M} \beta(a)\,\beta(b)\, D_{a,b} \cos\left[\frac{(2x+1)\,a\pi}{2M}\right] \cos\left[\frac{(2y+1)\,b\pi}{2M}\right] \tag{6.4}$$

here, $i_{x,y}$ represents spatial domain image coefficients.

6.2.3 Semantic structural constraint branch

As shown in Figure 6.1, sub-network2 (SSC-B) is an auxiliary branch guiding our PFH-B to generate images with three dimensional parametric feature information. This branch consists of consecutive five HFE-Bs followed by a reconstruction module to achieve image size same as of the final output image.



HR Face Image DCT of HR Face Image

Figure 6.4: Figure shows (from left to right) ground truth face image and corresponding discrete cosine transform

To add structural semantic constraint in the proposed face super resolution, we use 3D facial parameters fitted on a 2D image. In the following subsections, we have explained the surrey face model and procedure to get the target 2D image fitted with 3D parametric information obtained from surrey face model.

6.2.3.1 Surrey face model

High resolution 3D face scans obtained from the 3dMDface camera were utilized to build a 3D morphable model (surrey face model) [173] by bringing them in dense correspondence using an iterative multi-resolution dense 3D registration technique (IMDR) [174]. Resultant mesh constitutes two vectors: 1. shape vector represented by $\bar{S}\epsilon\mathbb{M}^{3L}$ 2. texture vector represented by $\bar{T}\epsilon\mathbb{M}^{3L}$. Here, $\bar{S}$ vector carries the coordinate information i.e. x, y and z components while per-vertex RGB data is present in $\bar{T}$ vector. Number of vertices used to built a mesh is represented by L. Principle components of both vectors ($\bar{S}, \bar{T}$) are extracted. The resulting 3DMM accounts for PCA shape model and PCA color model. Each model is mathematically represented by eq 6.5:

$$P := (\bar{m}, \sigma, M) \tag{6.5}$$

here, $\bar{m}\epsilon\mathbb{M}^{3L}$ represents example meshes mean. $M = [m_1, m_2, - -m_j]\epsilon\mathbb{M}^{3L\times j}$, j and M

represents number and set of principal components; and, σ is standard deviation which belongs to $\mathbb{M}^j$ to keep 99% of variance in ground-truth data. Also $j \leq u-1$, where, number of 3D scans, utilized to construct a 3DMM, are represented by u.

Novel faces using PCA shape model are generated by using equation 6.6

$$\bar{S} = \bar{m} + \sum_{i=1}^{j} \gamma_i \sigma_i m_i \tag{6.6}$$

here, shape coefficients are denoted by γ. These coefficients form the set constituting the instance coordinates of three dimensional face in PCA shape model.

The primary components of the PCA shape model signifies the shape of the face (slim, chubby or round etc). And the secondary components account for the more finer details (wrinkles, face irregularities, face depth etc) of face images. Whereas the primary component of PCA color model accounts for the change of face color from black to white and secondary components are related to gender. 3DMM shape model is sufficient to recover high facial details from a low resolution image. Hence, we considered only 3DMM shape model for our work.

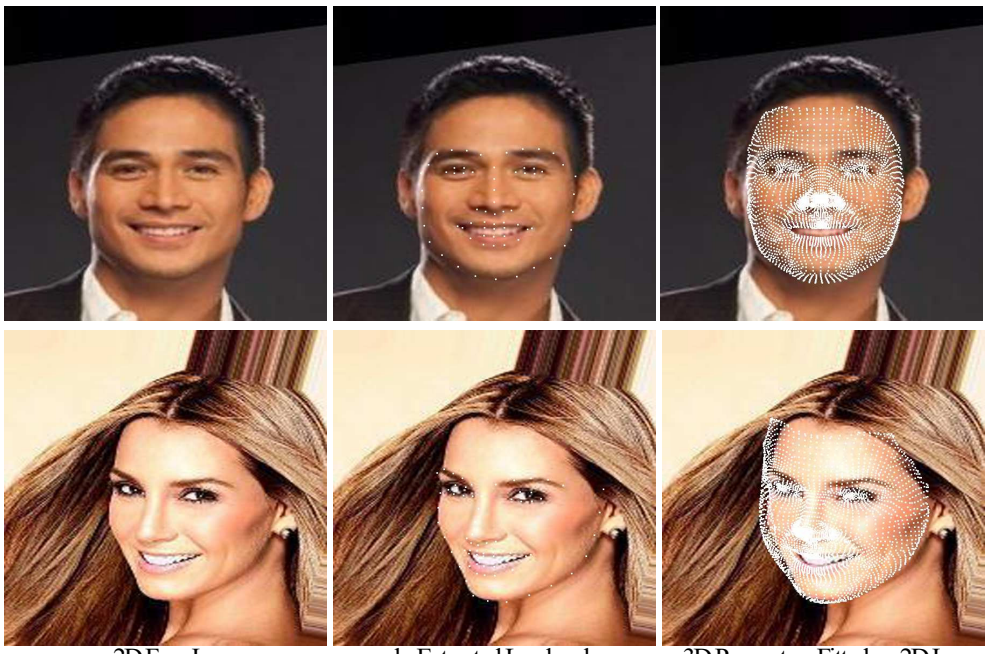
6.2.3.2 3D model fitting to 2D images

To obtain this image following steps are followed:

1. Firstly, facial landmarks of 2D facial image are extracted through dlib library [175].
2. We used PCA shape model of Surrey Face 3D Morphable model to obtain a 3D face mesh (refer eq 6.6).
3. Using EOS library, the face mesh obtained above is fitted to the extracted landmarks [170]. Four steps are followed to perform this shape-to-landmark fitting: Estimate the pose of the facial image, shape-specific identity fitting, linear expression fitting, and contour (which includes front facial contour and occluding contour) fitting [173].
4. The last step is to render the obtained 3D face mesh parameters ($v = [x, y, z]^t$) as 2D points ($v' = [x', y']^t$) on the face image (refer Figure 6.5). And for this translation, scaled orthographic projection [176] is used.

6.2.4 Loss Functions

To obtain the final output face image, loss function L^{fh} is optimized over M training samples. Loss Function L^{fh} is weighted combination of loss functions explained in this section.



a. 2D Face Image b. Extracted Landmarks c. 3D Parameters Fitted on 2D Image

Figure 6.5: Generation of training data for SSC-B. Figure shows (from left to right) a ground-truth 2D image, landmarks extraction on 2D images and then 3D parameters fitting on 2D image.

6.2.4.1 DCT based loss function

For subnetwork-1 i.e. autoencoder, L_1 loss between the generated coefficients from subnetwork-1 and HR DCT coefficients of high resolution face image is calculated. This loss function as this loss function is trying to penalizes the proposed network for not predicting the high frequency details correctly. The proposed DCT based loss function ($L^{fh}_{dct/i.j}$) is defined in eq. (6.7)

$$L^{fh}_{dct/i.j} = \frac{1}{W_{i,j}H_{i,j}} \sum_{a=1}^{W_{i,j}} \sum_{b=1}^{H_{i,j}} | \ \Delta_{i,j}(i^{hr})_{a,b} - (SN^1_{\Theta_{g1}}(\Delta_{i,j}(i^{lr})))_{a,b} \tag{6.7}$$

where, W and H represents the dimensions of DCT based HR coefficients. $\Delta_{i,j}(i^{hr})$ and $\Delta_{i,j}(i^{lr})$ are the DCT coefficients extracted from the ground truth HR face image and low resolution face image, respectively. $SN^1_{\Theta_g}$represents the subnetwork-1 and its parameters.

6.2.4.2 Semantic Structural Constraint Loss

To add three dimensional parametric information to obtain the final output image, we proposes semantic structural constraint loss. L_1 loss is calculated between the generated image from

the subnetwork-2 and its corresponding ground-truth image (2D HR face image fitted with 3D parametric information) and is given by eq. 6.8

$$L^{fh}_{ssc/i.j} = \frac{1}{W_{i,j}H_{i,j}} \sum_{a=1}^{W_{i,j}} \sum_{b=1}^{H_{i,j}} | \varsigma_{i,j}(i^{hr})_{a,b} - r (SN^2_{\Theta_{g2}}(i^{lr}))_{a,b} \quad (6.8)$$

here, $\varsigma_{i,j}(i^{hr})$ represents the 2D HR face image fitted with 3D parametric information. i^{lr} is the low resolution face image which is passed to subnetwork-2 ($SN^2_{\Theta_{g2}}$) to update its parameters.

6.2.4.3 Final loss function

Final loss function is the combination of dct based loss function, semantic structural constraint loss, feature based L_2 loss and adversarial loss. Feature based loss L^{fh}_{vgg} and adversarial losses L^{fh}_{adv} are explained it detail by Ledig et al. [69]. So, final loss function (L^{fh}) is the combination of four loss functions as mentioned in eq. 6.9

$$L^{fh} = L^{fh}_{vgg} + \alpha L^{fh}_{adv} + \beta L^{fh}_{dct} + \gamma L^{fh}_{ssc} \quad (6.9)$$

here, α, β and γ are the weight parameters used to balance the impact of individual loss functions.

6.3 Experiments

6.3.1 Datasets

From CelebA dataset [177], $108,640$ images are selected for the training purpose, 5000 images for validation and 5000 for testing purpose. To validate the performance of proposed architecture, we have performed experiments on benchmark face hallucination datasets- Menpo dataset (left, right and semi-frontal profile) [178] and Helen dataset [179].

6.3.2 Implementation details

The performance evaluation for the proposed architecture is performed for upscaling factors of $\times 4$ and $\times 8$ between the HR face image and LR face image. For an upscaling factor of $\times 4$, the

HR images are of size 128×128. These images are downsampled using bicubic kernel with a factor of $\times 4$ to generate a LR face images with size 32×32. For an upscaling factor of $\times 8$, the HR face images (128×128) are downsampled with a factor of $\times 8$ to generate low resolution images (16×16).

For every stage in progressive face hallucination branch, five HFE-Ms are used with fixed channel size (64) followed by a CEC-AM. In CEC-AM, the number of input channels are 256 and number of output channels are 64 with an expansion factor of 3. To reduce the number of parameters in FC layer, total channels present is a layer are divided by a factor R having value 24. The LeakyReLU hyper parameter α is 0.2. Batch size is 8; optimizer used is Adam with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Initial learning rate $= 0.0001$. To minimize the distance between the HR face image and generated face image, l^{fh} 6.9 is used. To fully train the model, alternate training between the generator and discriminator is done to update their weights. Proposed model is evaluated using two commonly used metrics: SSIM and PSNR. For equitable comparison with the previous works, these metrics are computed on Y-channel.

6.3.3 Ablation study

In order to understand the importance of individual sub-modules of the proposed architecture, ablation study is conducted as summarized in table 6.1.

Firstly, we studied the effect of using upsampling layer at different positions in the architecture i.e. a single stage upsampling model vs a progressive stage upsampling model. From the results obtained after employing upsampling layer at end of the architecture (refer Figure 6.6a) and at progressive stages (refer Figure 6.6b), it is clear that multi-stage upsampling performs better than single stage upsampling model. As progressive upscaling approach allows the network to mimic the fine details present in an input image and increases its ability to learn.

The proposed architecture is compared with or without using CEC-AM. There is significant improvement in the results after adding this module in the architecture (refer Figure 6.6b and 6.6c). This module is basically used to slant the network's ability to provide access of available resources to the most important information present in the feature maps.

In order to further improve the quality of the generated image, we embed sub-network1 in the architecture (refer Table 6.1d). Results obtained (refer Figure 6.6d) after adding sub-network1 in the proposed architecture supports our claim that DCT based auto encoder is able to add high frequency information in the architecture.

Still some facial details like skin irregularities and depth information is missing in the generated image. So, we tried to add these facial details using sub-network2 (refer Table 6.1e). In this experiment we used the progressive upscaling and sub-network2 and excluded the sub-network1. Results obtained shows substantial improvement in the generated images perceptually. As facial

Table 6.1: Contribution of different components utilized in the proposed architecture

Components	a	b	c	d	e	f
Single stage	✓					
Multiple stages		✓	✓	✓	✓	✓
HFE-M	✓	✓	✓	✓	✓	✓
CEC-AM			✓	✓	✓	✓
Sub-network1				✓		✓
Sub-network2					✓	✓

details and skin irregularities are more prominent in these images. But there is little mismatch in the color as compare to the ground-truth images (refer Figures 6.6e and 6.6g). So, for the final architecture we combined both the approaches i.e. DCT based auto encoder and 2D images with 3D parametric information to get the final output images (refer Figure 6.6f), performs better quantitatively and qualitatively.

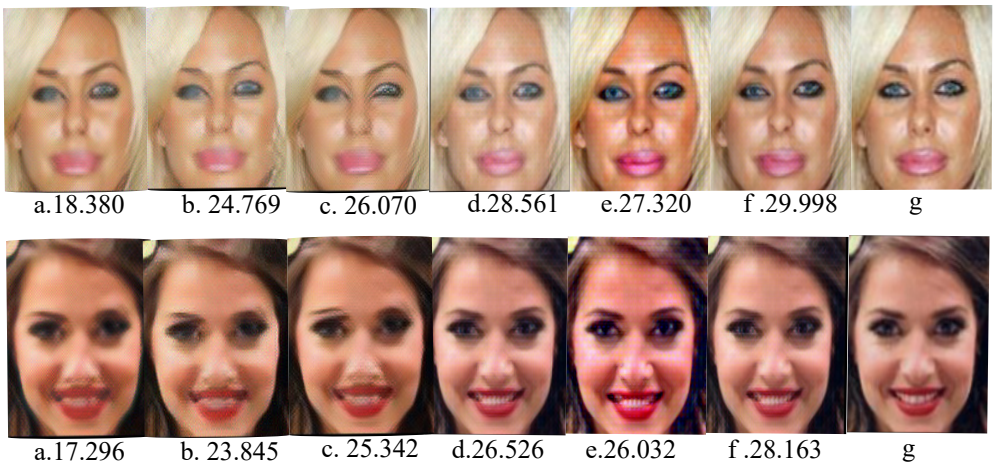


Figure 6.6: Investigation of different components utilized in the proposed architecture: a) Single stage upscaling with HFE-M in generator, b) Multiscale upscaling with HFE-M, c) Adding Channel attention mechanism in b. d) Using sub-network1 (DCT based autoencoder) along with c. e) Using sub-network2 (Semantic Structural constraint block) along with c. f) Using both sub-network1 and sub-network2 along with c.

6.3.4 Comparison with the SOTA method

We compared our proposed architecture with seven SOTA methods: SRCNN [144], VDSR [66], SRGAN [69], ESRGAN [84], ImprovedFSR [180], SICNN [181], SAM3D [182] and bicubic interpolation to show the efficacy of our network. For fair comparison, we trained all these models on our training dataset with same parameters.

Menpo dataset- We evaluated the performance of our model and other SOTA methods on Menpo dataset (left, right and semi-frontal profiles) [178] qualitatively and quantitatively. For

left profile, our model has achieved second highest PSNR and highest SSIM for both $\times 4$ and $\times 8$ scaling factors (refer Table 6.2). For right profile our model has achieved highest PSNR and SSIM for $\times 4$ scale. For $\times 8$ scale, our model has achieved second highest PSNR and highest SSIM values. For semi-frontal profile $\times 4$ scale, our model has second highest PSNR and highest SSIM values and highest PSNR and SSIM numbers for $\times 8$ scale.

Perceptual results for Menpo dataset are represented in figures 6.7, 6.8 and 6.9. In figure 6.7, although the image generated by ImprovedFSR has highest PSNR, still there are artifacts present in the eyes and mouth region. In SAM3D model, the generated image has blemished skin with noise present in the mouth region. Only the image generated by proposed model is able to mimic the ground-truth image. As shown in figure 6.8, the image generated by SRGAN model has some artifacts in the hair region. And none of the competing GAN methods are able to generate a mole present on the face. Our method is recovering small details or textures present in the face image. In figure 6.9, competing GAN methods are unable to recover the details of the nose region. And CNN based methods are producing smooth faces with very less textural details. Image generated by our model is looking perceptually better than all other methods.

Table 6.2: Quantitative result comparison on the basis of average PSNR (dB) and average SSIM on different facial poses (left, right and semi-frontal) of Menpo dataset.

Scale	$\times 4$						$\times 8$					
	Left		Right		Semi-frontal		Left		Right		Semi-frontal	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Bicubic	27.32	0.794	26.28	0.772	24.89	0.763	22.01	0.634	21.07	0.602	20.31	0.501
SRCNN[144]	26.99	0.811	26.59	0.794	25.15	0.772	22.14	0.663	21.16	0.623	20.99	0.509
VDSR[66]	27.45	0.831	26.78	0.793	25.32	0.769	22.56	0.671	21.12	0.625	20.61	0.512
SRGAN[69]	29.34	0.873	29.51	0.871	28.94	0.868	22.34	0.681	22.31	0.661	21.02	0.557
ESRGAN[84]	28.99	0.821	28.83	0.813	28.43	0.813	21.04	0.662	21.06	0.621	20.19	0.532
SICNN[181]	28.09	0.812	28.12	0.803	27.02	0.799	22.45	0.679	22.69	0.659	21.21	0.613
ImprovedFSR[180]	**30.85**	0.887	30.78	0.884	**30.25**	0.877	**23.90**	0.713	**23.56**	0.676	21.28	0.623
SAM3D[182]	28.23	0.842	28.91	0.848	27.92	0.838	23.71	0.692	23.01	0.654	21.11	0.601
Ours	30.36	**0.924**	**30.81**	**0.925**	29.80	**0.922**	23.73	**0.723**	23.52	**0.689**	**21.34**	**0.641**

Helen dataset- For Helen test dataset [179], quantitative results (PSNR/SSIM) are shown in Table 6.3 for an upscaling factor of $\times 4$. Our method is able to achieve hightest PSNR and SSIM values as compare to SOTA methods. Perceptual analysis with the PSNR and SSIM values is presented in figure 6.10. CNN based methods like SRCNN and VDSR are generating output images with blurriness. SRGAN and ESRGAN are also producing images with some artifacts at the eyes and cheeks. Other competing methods like ImprovedFSR and SAM3D are unable to generate better images perceptually. Only the proposed method is able to recover fine textural details like eye brows and depth details similar to the ground-truth image.

Proposed model has achieved highest SSIM and second highest PSNR for $\times 8$ factor (refer Table

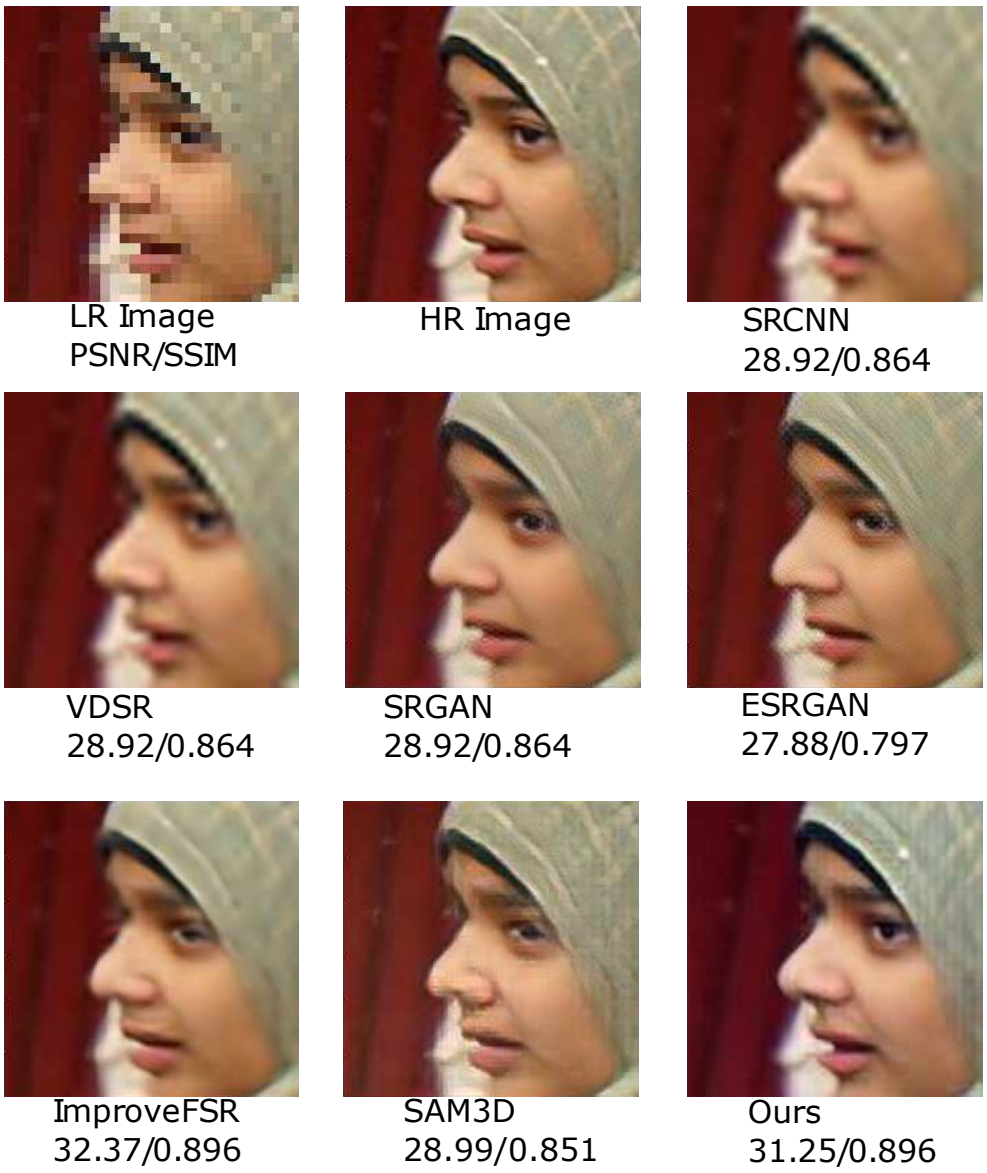


Figure 6.7: Perceptual and quantitative (PSNR/SSIM) result comparison with SOTA methods for magnification factor of $\times 4$ on Menpo test dataset.

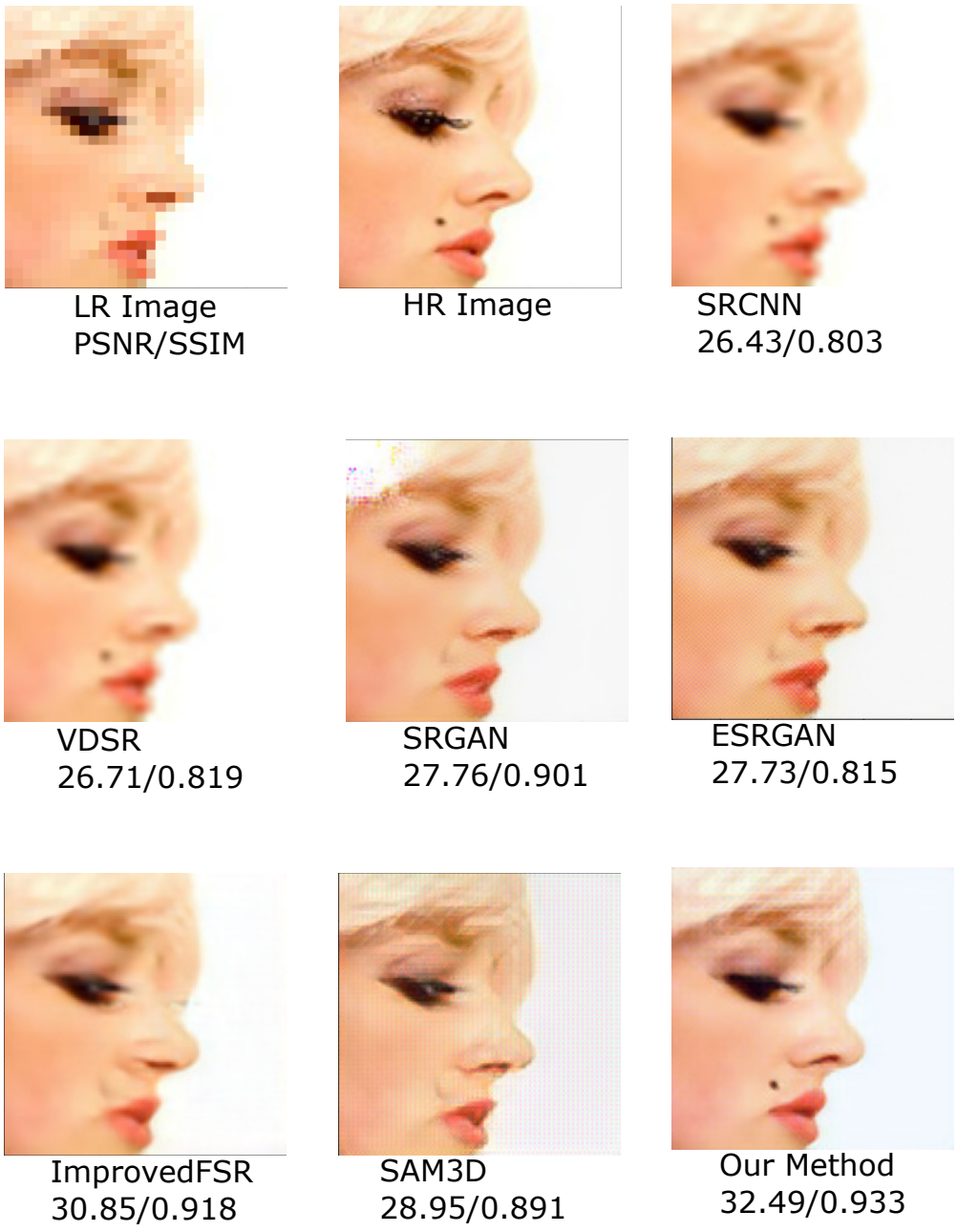


Figure 6.8: Perceptual and quantitative (PSNR/SSIM) result comparison with SOTA methods for magnification factor of $\times 4$ on Menpo test dataset.

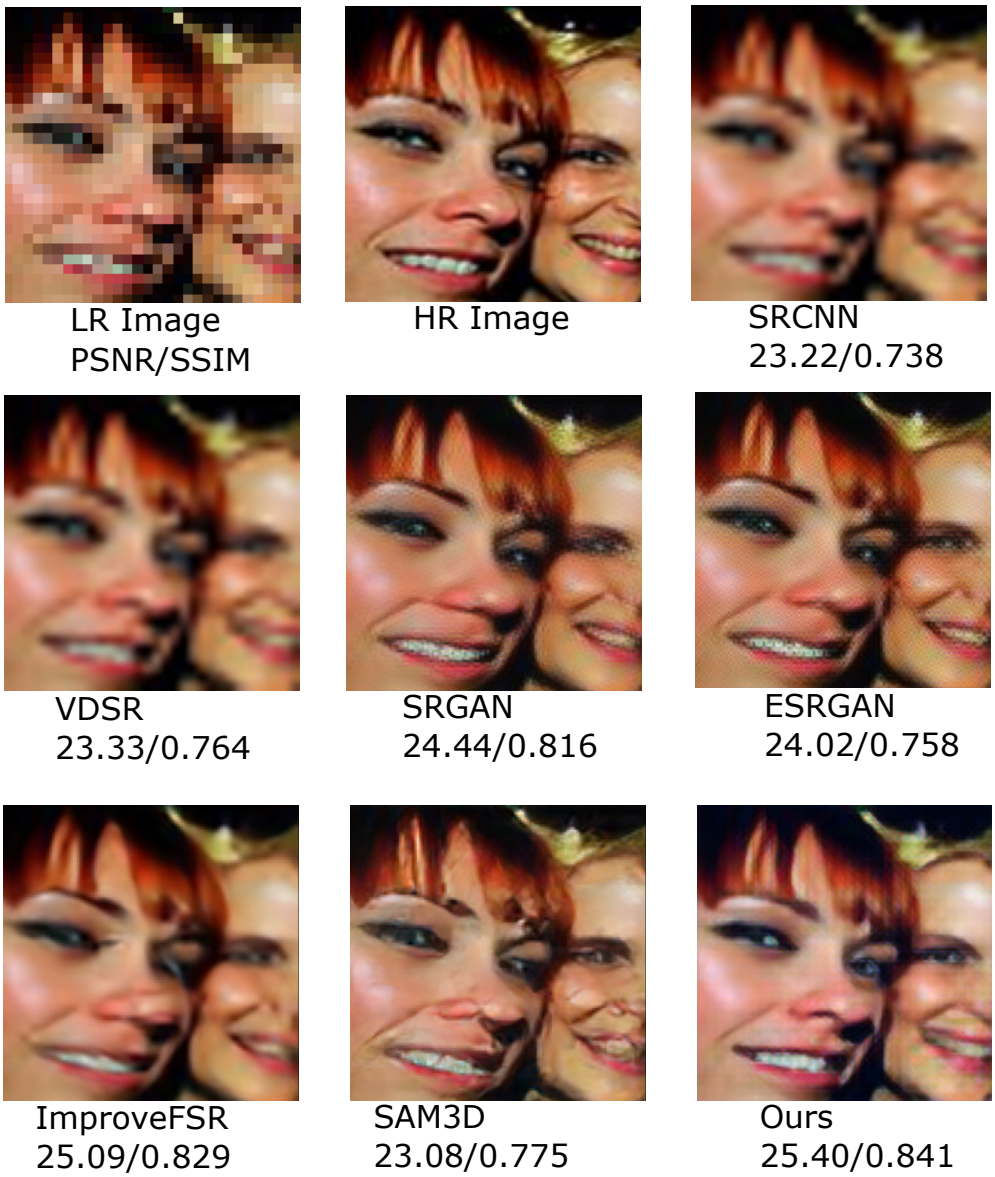


Figure 6.9: Perceptual and quantitative (PSNR/SSIM) result comparison with SOTA methods for magnification factor of $\times 4$ on Menpo test dataset.

Table 6.3: Quantitative result comparison on the basis of average PSNR (dB) and average SSIM of Helen dataset.

Scale	×4		×8	
	PSNR	SSIM	PSNR	SSIM
Bicubic	25.06	0.692	21.67	0.612
SRCNN[144]	26.45	0.712	22.04	0.634
VDSR[66]	26.89	0.734	22.14	0.639
SRGAN[69]	28.45	0.851	22.31	0.689
ESRGAN[84]	27.86	0.790	21.99	0.674
SICNN[181]	26.43	0.757	22.76	0.681
ImprovedFSR[180]	28.83	0.856	**23.99**	0.701
SAM3D[182]	27.32	0.834	22.16	0.712
Ours	**28.86**	**0.911**	23.83	**0.741**

6.3). Qualitative analysis with their quantitative numbers are presented in figures 6.11 and 6.12. From the figures, it is clear that perceptually our model is performing better than the other competing methods. ImprovedFSR has highest PSNR value but visually our model is able to recover more finer details and textures.

6.4 Conclusion

Current work presents a novel GAN based progressive face hallucination network. To generate the final output image with 3D parametric information, proposed model uses a auxiliary supervision network which is compelled to generate 2D images with 3D parametric information using shape model of 3DMM. To incorporate high frequency components in the image, an AE is proposed which generates HR coefficients of DCT. To embed HR DCT information into the face hallucination network IDCT block is introduced within the network to convert the frequency domain coefficients to spatial domain. Output images generated by the proposed model have subtle structural details with depth information, outperforming the SOTA methods.

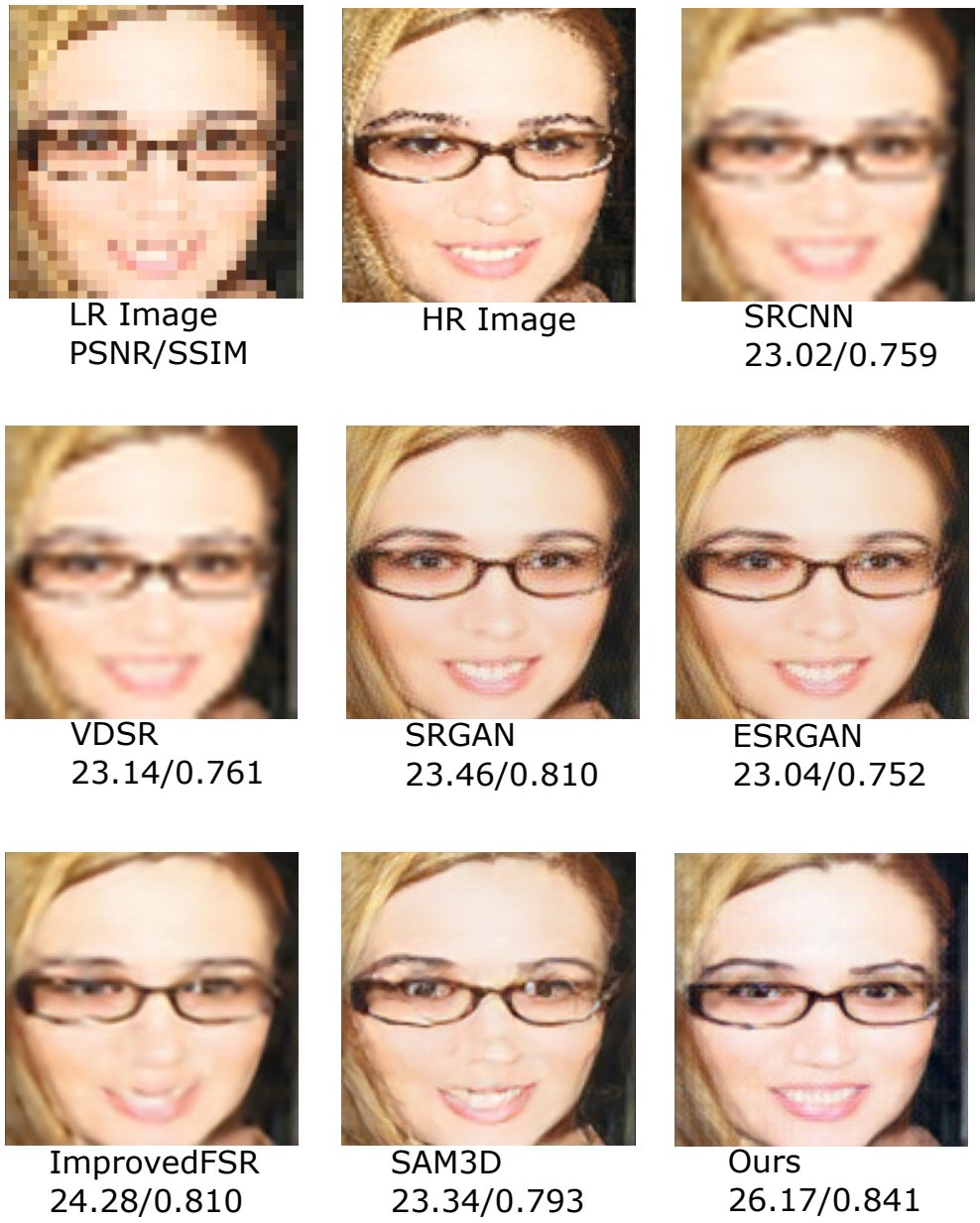


Figure 6.10: Perceptual and quantitative (PSNR/SSIM) result comparison with SOTA methods for magnification factor of $\times 4$ on Helen test dataset.

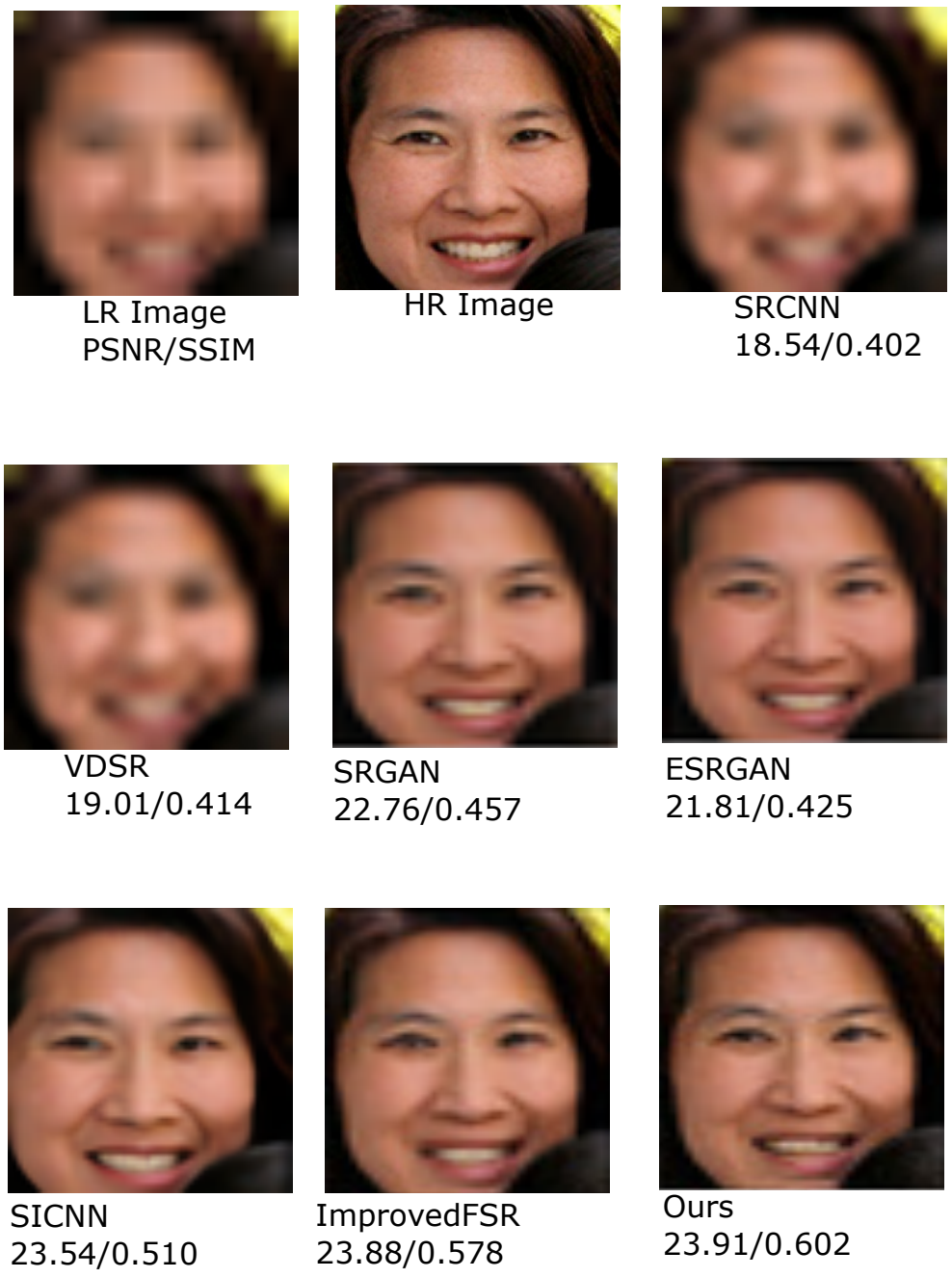


Figure 6.11: Perceptual and quantitative (PSNR/SSIM) result comparison with SOTA methods for magnification factor of $\times 8$ on Helen test dataset.

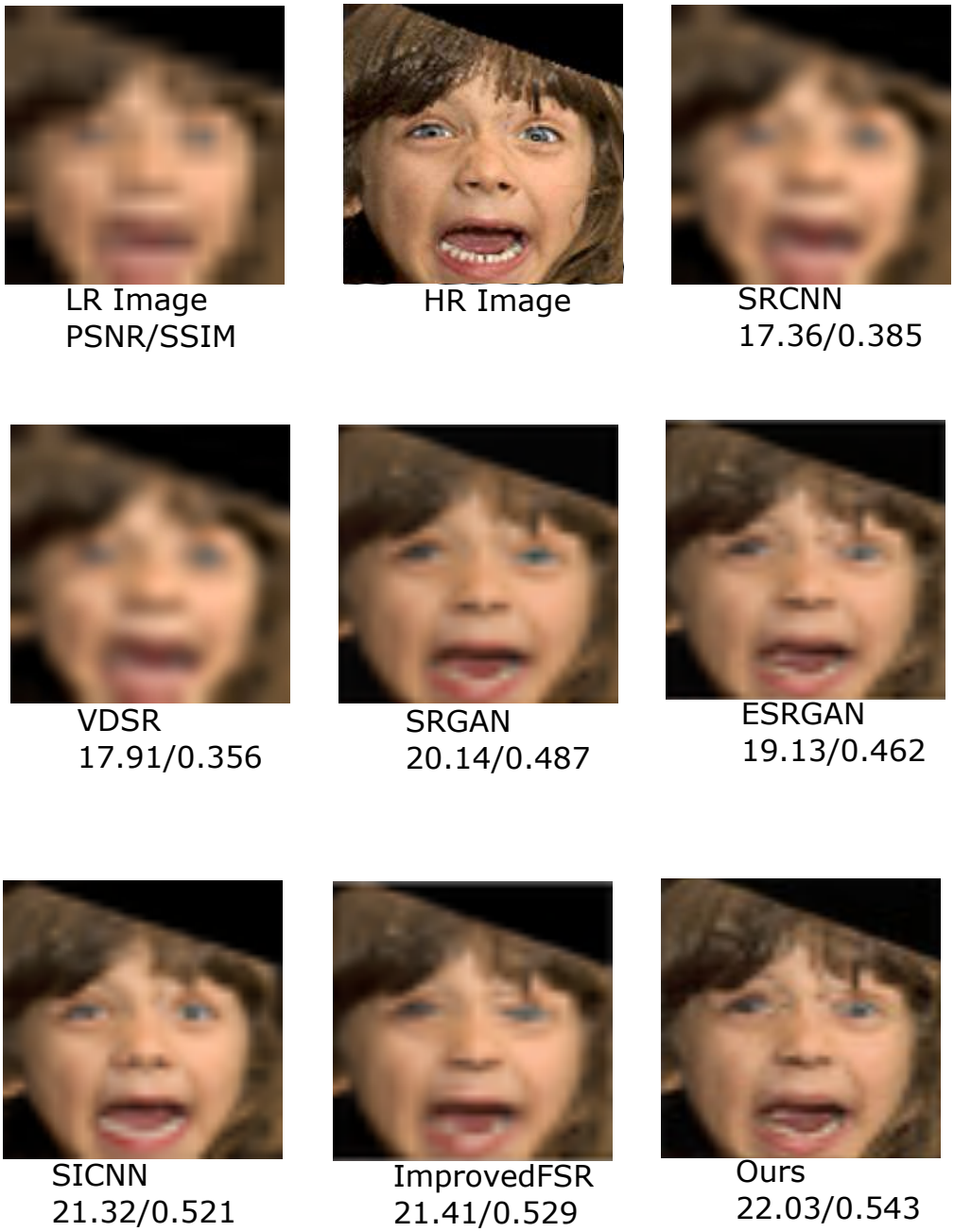


Figure 6.12: Perceptual and quantitative (PSNR/SSIM) result comparison with SOTA methods for magnification factor of $\times 8$ on Helen test dataset.

7 Video face hallucination with frequency supervision and cross modality support

7.1 Introduction

Resolution enhancement of the facial images has a wide range of applications in different fields of computer vision, such as face recognition [183], deep fake generation and detection [184, 185], emotion detection [186], face identification [187] and face alignment [188], etc.

Recently, a number of approaches have been proposed to solve the problem of face hallucination in static images [182, 189, 190]. Despite these breakthroughs, there are still numerous challenges in Video Face Hallucination (VFH). The temporal and spatial information are two critical prerequisites for VFH models. Since, only a single image is considered in current face hallucination models - the spatial information has been predominantly used in most of the SOTA methods. Whereas, to solve the VFH problem, temporal motion of the spatial features is a key component. Some recent works aim to incorporate this temporal information in VFH models, for example, [136, 137] uses 3D convolution network to learn temporal consistency in the network for VFH.

However, these methods have very high computational resource requirements. The other type of methods [191, 135] use stacks of multiple low resolution frames from videos and pass them to the deep neural network. Since, these methods naively fuse multiple frames at the input, no implicit leaning mechanism is used to learn the temporal relationship between the spatial features. The semantic priors (such as, facial parsing maps) and 3D priors also assists the VFH models to generate superior results. Yet, the requirement of pre-trained networks to obtain a better prior leads to extra computation cost.

One major limitation of generative models is their inability to learn the whole range of frequencies present in a pattern [192]. These models prioritise a smaller band of frequencies instead of learning the full spectrum. Hence, the salient local features of the regions are harder to learn as they are generally absent in low resolution images. This issue has been overlooked in the GAN

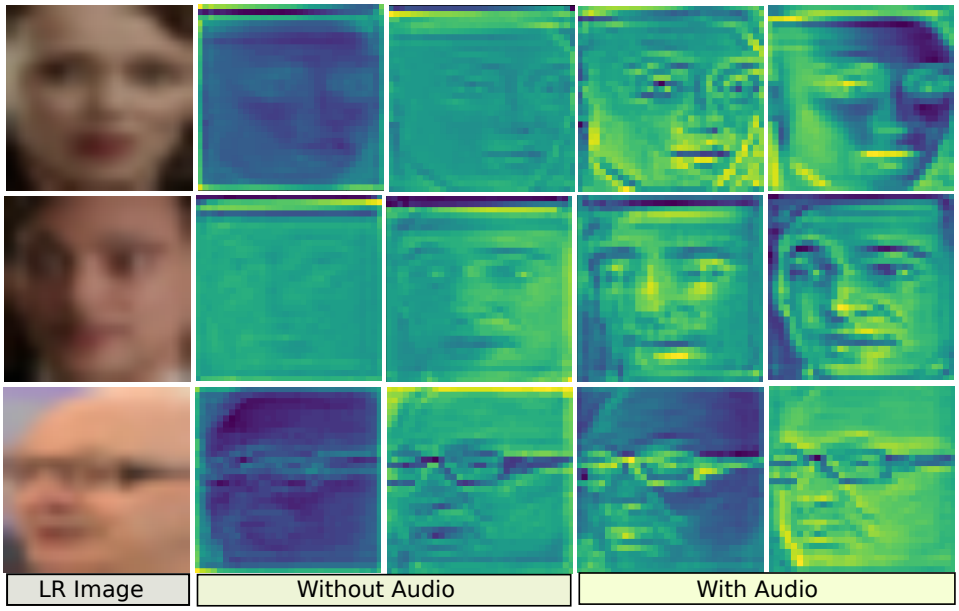


Figure 7.1: Visualisation of activation maps from the generator network to show the importance of attention based audio network.

based face super-resolution literature. Most of the SOTA use objective function that only optimize the spatial domain features. Consequently, learned feature representations do not contain the high frequency detail leading to lower sharpness of the visual output. Furthermore, current VFH approaches have blurriness in regions with larger motion, such as mouth, lips, teeth, etc. The root cause of this problem is absence on implicit mechanism to learn temporal dependence between the frame. Higher structural complexity of a region is also partially responsible for degradation of quality.

In is a known fact that audio signal and the motion in facial video have very high correlation [193]. This correlation can be used to implicitly learn the mechanism for motion in different facial regions. However, this has never been explored up to now. We hypothesis that an objective function capturing the speech variation can also build the temporal consistency in the visual outputs. Furthermore, speech signal carries a lot of information regarding gender, identity and age [112]. Hence, audio signal can play a critical role in re-identification of such information, especially in the very low resolution images where gender and identity details are almost lost (refer Figures 7.1 and 7.3).

In this chapter, we exploit the correlation between the frequency spectrum of the speech signal and the motion of spatial regions (like, mouth and lips). We present a multi-modal GAN architecture that uses both speech and video modalities to further enhance the quality of SR for facial images. The proposed method uses two feature encoder in the generator. First encoder extracts the spatial features from the facial image while second extracts the audio features at corresponding time step. An axial self attention mechanism [194] is used for the feature fusion between both the modalities.

We also use a frequency based loss function (derived from 2D discrete Fourier transform) in order to learn high frequency details which are otherwise difficult to generate. Furthermore, we present a lip reading loss to resolve the issue of blurriness in large motion regions. Instead of directly minimizing the distance between generated image and the ground-truth image - distance between their corresponding feature maps (extracted from the intermediate layers of the pre-trained lip-reading network) are minimized. The proposed loss function propels the Video Face Hallucination Network (VFHN) to generate faces with very sharp mouth region. Main contributions are as follows:

- We proposed first multi-modal architecture with cross-modality support to learn dependence between audio and facial videos to generate fine grained spatial-temporal motion and retrieve facial identity information.
- An explicitly defined lip reading loss is presented to further improve the temporal consistency and remove the blurriness in key facial regions.
- A Fourier transform based frequency domain loss is also applied to add the salient frequency features.
- Visual results, quantitative numbers along with their edge restoration number (metric used to examine the restoration quality of edges in videos) show superiority of proposed work over other SOTA methods.

7.2 Proposed Methodology

Frequently, videos that capture faces, such as those recorded by webcams or mobile cameras, include accompanying audio. However, the VFH literature has not explored the semantic correlation between the audio features and facial frames. We introduce a novel multi-modal architecture with cross-modality support to investigate the significance of aural features in facial videos. To further enhance the quality of our outputs, we incorporate a Fourier transform based frequency domain loss to highlight important frequency features. Additionally, we address issues with blurriness in regions such as the mouth and chin and ensure temporal consistency across consecutive video frames by introducing a lip reading-based loss function. This loss function encourages our VFH network to produce outputs with improved texture details around the mouth region while maintaining consistency across frames. This section provides a comprehensive overview of our proposed architecture and the various loss functions that were developed for it.

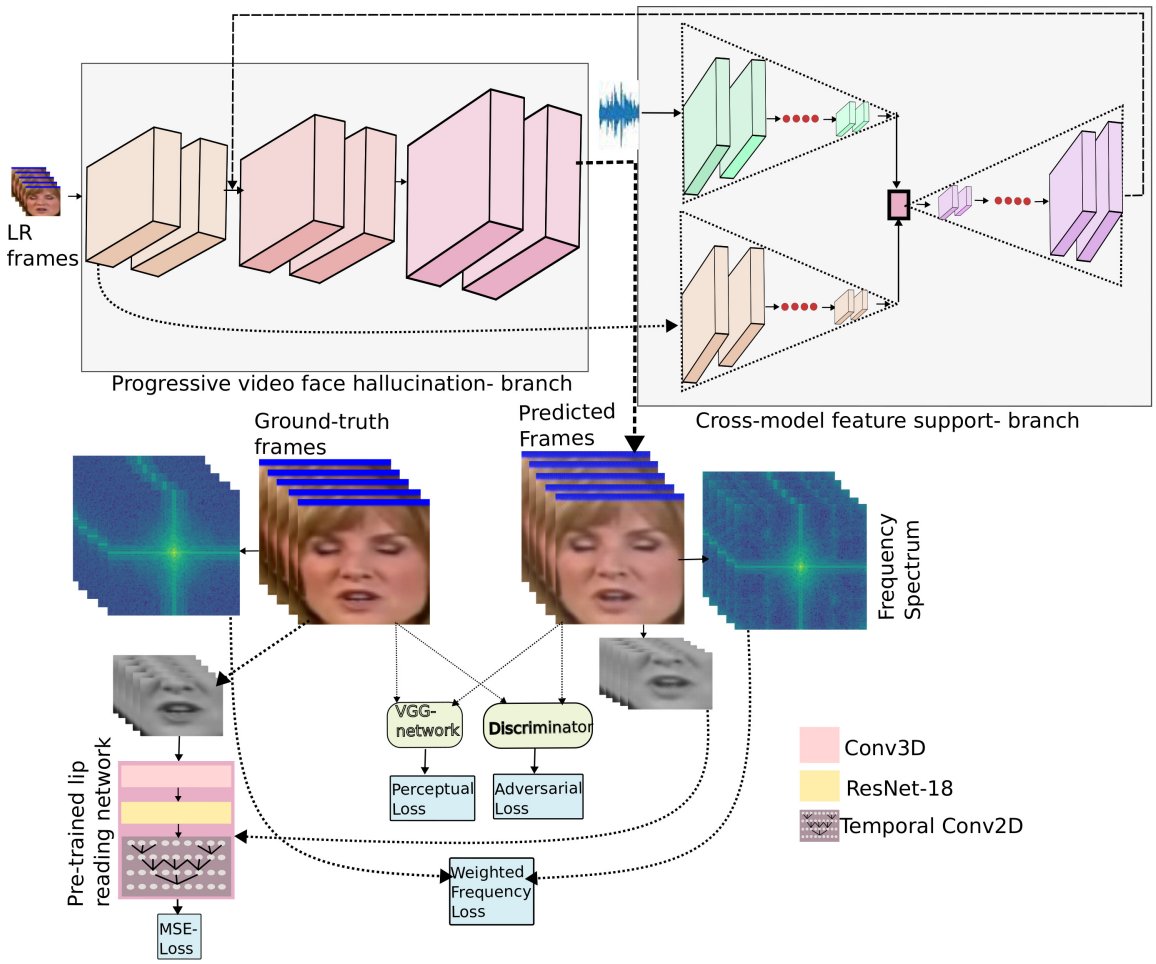


Figure 7.2: VFHN: final output is generated by progressive hierarchical feature extraction branch (PHFE-B). Audio features are extracted from the cross-modal feature support- branch (CMFS-B). Since, PHFE-B is based on progressive upscaling, so aural embedding are merged at the very initial stage of the network to add semantic supervision.

7.2.1 Overview

The aim of the proposed video face hallucination network (VFHN) is to generate a high resolution video frame (f_t^{hr}) from corresponding low resolution frame (f_t^{lr}), its temporal neighborhood (f_{t-j}^{lr},... f_{t-1}^{lr},f_{t+1}^{lr},..f_{t+j}^{lr}) and audio signal (a_t) (refer eq 7.1). Here, t refers to the time step at which face hallucination is performed and $2j$ is the size of temporal neighborhood window used to support the super-resolution. VFHN (ζ) further comprises of two branches. The first branch is the progressive hierarchical feature extraction branch (PHFE-B), which generates a high-resolution video through a series of hierarchical feature extraction steps. The second branch is the cross-modal feature support branch (CMFS-B), which employs an auto-encoder based architecture to extract features from the audio signals. The output feature maps obtained from the second module are merged in the PHFE-B, guiding it to generate facial videos with correct identity and gender. These two modules are explained in detail, later in this section.

Since, CNNs are predominantly used in image data due to their established superiority in modeling spatial context. Furthermore, short instantaneous audio (near to the input frame) does not provide very useful information due to lack of context. Therefore, rather than directly applying audio signal to CMFS-B, we transformed it into frequency representation (spectrum) using short time Fourier transform: $a_f = STFT(a_T)$ (here, T represent full length of sequence). This spectrum is further converted into mel-spectograms (a_{m_s}) for better perception and used as an input to our CMFS-B (refer eq 7.1).

$$f_t^{sr} = \zeta((f_t^{lr}, \{f_i^{lr}\}_{i=t-j}^{t+j}), a_{m\_s}; \emptyset_{w_v,b_v}, \emptyset_{w_a,b_a}) \tag{7.1}$$

here, f_t^{sr} is the generated high-resolution output frame and ζ represents the VFHN. $\emptyset_{w_v,b_v}$ and $\emptyset_{w_a,b_a}$are the weights and bias parameters of video processing module (PHFE-B) and audio processing module (CMFS-B), respectively. For optimization of both the modules, multiple loss functions (for each independent modality and joint for cross-modal) are used to control learning of parameter.

Proposed architecture (VFHN) is depicted in Figure 7.2. The motivation for the visual branch (PHFE-B) of VFHN is taken from [195]. This branch comprises of two sub-modules. The first module rely on the complex filter formation, where input feature maps are passed through different convolution layers with increasing filter sizes ($1 \times 1, 3 \times 3, 5 \times 5, 7 \times 7$). Different size kernels allow the network to assimilate primary (local) attributes as well as hierarchical (global) features. Second module is channel attention module where depth-wise convolution layers [196] are utilized. These layer compute association on short stacks of feature making them computationally efficient. Further, we used a squeeze and excitation block [172] on the output of previous block to encourage the network to emphasize on extracting feature maps from channels which are more affluent.

As shown in Figure 7.2, CMFS-B module also contain two further sub-modules. First sub-module takes feature maps from the convolution layer of the first stage of of PHFE-B as input ($\xi(f_t^{lr'}; \emptyset_{w_{v'},b_{v'}})$) while audio mel-spectogram is fed as input to the second sub-module ($\psi(a_{m\_s}; \emptyset_{w_a,b_a})$).

Feature embedding obtained from these two sub-modules are concatenated and are further converted into a latent vector ($f^{lv'}$) (refer eq. 7.2). This latent vector f^a (refer eq. 7.3) is applied to axial attention layer [194] to capture the long range dependencies and identify important features with higher relevance.

$$f^{lv'} = \Psi\{(\xi(f_t^{lr'}; \emptyset_{w_{v'},b_{v'}}) + \psi(a_{m\_s}; \emptyset_{w_a,b_a})); \emptyset_{w_{lv},b_{lv}}\} \tag{7.2}$$

$$f^a = axial\_atten(f^{lv'}) \tag{7.3}$$

Obtained feature maps (f^a) are applied to a series of hierarchical feature extraction steps followed an upscaling using a sub-pixel layer [68]. Since, the proposed architecture follows the progressive upscaling, hence at each stage the feature maps are upscaled by a factor of 2. The final output video facial frames (f_t^{sr}) generated by the proposed model are the enhanced and upscaled version of low resolution input video facial frames with correct identity, gender information and temporal consistency across the frames.

7.2.2 Loss function

The overall loss function, used to optimise whole GAN architecture, is weighted combination of five loss functions, which are: lip-reading based loss ($L^{lf_1}_{L\_Rd/l.m}$), weighted frequency loss ($L^{lf_2^*}_{freq}$), perceptual loss ($L^{lf_3}_{vgg}$) [69], adversarial loss ($L^{lf_4}_{adv}$) [182] and L1 Loss [115], (refer eq. 7.4).

$$L^{lf} = L^{lf_3}_{vgg} + \alpha L^{lf_1}_{L\_Rd/l.m} + \beta L^{lf_2^*}_{freq} + \gamma L^{lf_4}_{adv} + \delta L^{lf_4} \tag{7.4}$$

.

here, coefficients ($\alpha, \beta, \gamma and \delta$) are used to control the contribution of each loss. After empirical analysis, we found optimal values for best results to be: α= 0.0001, β= 0.001, γ= 0.0001 and δ= 0.01.

The output video frame (f_t^{sr}) is generated by optimizing the model for overall loss function (L^{lf}), calculated between the ground-truth video frame (f_t^{gt}) and generated video frame (f_t^{sr}) over N training data samples.

Perceptual loss is computed by comparing the feature representations of the generated high resolution image and the ground-truth image at a layer of pre-trained VGG network. Adversarial loss is computed as the binary cross-entropy loss between the discriminator's prediction of whether an image is generated or real and the true label. Lip reading loss and weighted frequency loss is explained in detail in the next subsection.

Lip reading loss function

Accurately identifying the lip movements is elemental for visual speech recognition. Hence, the lip-reading applications involve explicit learning mechanisms for movement of teeth, tongue

and lips. Another prerequisite for correct lip-reading is extraction of temporal dependency across consecutive video frames. These attributes of lip-reading networks can assist video face hallucination network in learning key scene components for generating finer texture of mouth region and maintaining temporal consistency across frames. We used pre-trained lip-reading network [197] to extract these texture rich feature embedding. The overview of the lip-reading model we used is as follows.

Firstly, a sequence of video frames with cropped mouth region are passed as input to the 3D convolution layer (refer eq. 7.5).

$$f^{sr'} = Conv3D(\{f_i^{sr}\}_{i=t-j}^{t+j}) \tag{7.5}$$

here, $\{f_i^{sr}\}_{t-j}^{t+j} \epsilon \mathbb{R}^{b*k*h*w}$, b is the batch size, k represents the number of frames in the video sequence, height and width of video frame is denoted by h and w, respectively. The $f^{sr'}$ is the output feature embedding.

These features $f^{sr'}$ are passed to ResNet-18 ($f^{sr''} = ResNet(f^{sr'})$) after reshaping them from $\boldsymbol{b*k*h*w}$ to $\boldsymbol{B*h*w}$. B is the final batch size, generated by concatenating number of frames (k) over batch dimension (b). The output visual features from Resnet-18 ($f^{sr''}$) are sent to multi-scale temporal convolution network with dilated convolution layers. Use of dilated convolution assists the network in learning long distance spatial contexts from the feature maps.

We used this lip-reading network to extract high level feature embedding (refer eq 7.6).

$$L_{L\_Rd/l.m}^{lf_1} = \frac{1}{w_{l,m}h_{l,m}} \sum_{x=1}^{w_{l,m}} \sum_{y=1}^{h_{l,m}} \aleph_{l,m}(\{f_i^{gt}\}_{i=t-j}^{t+j})_{x,y} - \aleph_{l,m}(\{f_i^{sr}\}_{i=t-j}^{t+j})_{x,y} \tag{7.6}$$

here, $L_{L\_Rd/l.m}^{lf_1}$ is the lip-reading loss. $\aleph_{l,m}(\{f_i^{gt}\}_{i=t-j}^{t+j})$ and $\aleph_{l,m}(\{f_i^{sr}\}_{i=t-j}^{t+j})$ are the feature embedding obtained from the intermediate layers of lip-reading network. L_1 loss is calculated between these feature embedding. The L_1 loss forces VFHN to generate images with better texture in key facial region and maintain temporal consistency across the sequence of videos.

Weighted Fourier Frequency Loss

Frequency representation of image gives better perception of artifacts present in an image [198]. Missing high frequency components can lead to ringing artifacts in the image. Whereas, sole presence of high frequency information will provide an image with just region boundaries and edges. Checkerboard artifacts can be seen in images when a band stop filtering is applied in

frequency domain [199]. From the above, it can be concluded that absence of some frequency components can lead to various artifacts in the spatial domain. Hence, we hypothesise that incorporating these missing frequencies in the frequency domain should lead to better perceptual quality in spatial domain. This has been validated by our results in table 7.1

We use 2D discrete fourier transform (DFT) to generate the frequency domain representation of an image (refer eq. 7.7, 7.8).

$$F(u,v) = \sum_{x=0}^{w-1}\sum_{y=0}^{h-1}(P_i(x,y))_{i=t-j}^{t+j}.(cos2\pi(\frac{ux}{w}+\frac{vy}{h})- isin2\pi(\frac{ux}{w}+\frac{vy}{h})) \tag{7.7}$$

$$F^{\star}(u,v) = \sum_{x=0}^{w-1}\sum_{y=0}^{h-1}(P_i^{\star}(x,y))_{i=t-j}^{t+j}.(cos2\pi(\frac{ux}{w}+\frac{vy}{h})- isin2\pi(\frac{ux}{w}+\frac{vy}{h})) \tag{7.8}$$

here, $P(x,y)$, $P_i^{\star}(x,y)$ are the pixel values of ground-truth and generated video frames at x and y coordinates, respectively.

Frequency spectrum coordinates are represented by u and v and its value by $F^{\star}(u,v)$. Mean square error is calculated between the ground-truth and generated image in the frequency domain as shown in eq. 7.9.

$$L_{freq}^{l2} = \frac{1}{wh}\sum_{u=0}^{w-1}\sum_{v=0}^{h-1}|F(u,v) - F^{\star}(u,v)|^2 \tag{7.9}$$

Generative models are more inclined towards generating easy (low) frequencies as compare to hard (high) frequencies [200, 201]. Since, each frequency value have same weightage and inherent biasing allows generative models to learn easy frequencies better than the hard frequencies. During training, to put more weightage to hard frequencies, a weight matrix (refer eq . 7.10), similar to the shape of spectrum, is introduced which adds non-uniformity to each frequency component in the cost function.

$$m(u,v) = |F(u,v) - F^{\star}(u,v)| \tag{7.10}$$

$m(u,v)$ is the weight matrix having range $[0,1]$, where weights near 0 signifies the frequencies with more weightage and 1 signifies the frequency which is getting vanished. Therefore,

frequencies which are learned by the model easily are down-weighted. And hence, the final weighted frequency cost function ($L_{freq}^{l2^*}$) is shown in eq. 7.11.

$$L_{freq}^{lf_2^*} = \frac{1}{wh}\sum_{u=0}^{w-1}\sum_{v=0}^{h-1}\sum m(u,v)||F(u,v) - F^{\star}(u,v)|| \tag{7.11}$$

7.3 Experiments

7.3.1 Datasets and Metrics

There are no publicly available datasets for video face hallucination which also contains corresponding audio signals. For training purpose, we selected LRW dataset [202] which is collection of videos having 500 words spoken in different sentences from more than 1000 speakers. We carefully selected videos from the dataset with maximum variance and divided them in training, validation and test sets. Facial landmarks estimated from OpenFace [203] are used to obtain a square crop and eliminate the undesired background from the frames. Cropped image is resized into fixed 128×128 size and then further downsampled to 32×32 for input LR image. For testing purpose, in addition to LRW dataset [202], we use grid speech corpus [204], VHFQ [205], Voxceleb [206] and LRS2 [193] datasets. All these datasets are audio-visual datasets, hence they satisfy the requirements for evaluation of proposed architecture.

In addition to the PSNR and SSIM (both calculated on Y channel after transforming RGB image into YCbCr), we also evaluate our results using edge quality assessment metric (ERQA). This metric is used to estimate the network's capability to restore the real information present in the videos.

7.3.2 Ablation study

Audio signal analysis: To validate the proposed hypothesis that the audio signal can play critical role in retrieving the lost visual data, we first train VFHN without the CMFS-B. The empirical (Table 7.1) as well as the visual results (Figure 7.3) verify that use of audio modality significantly improves the quality of the output SR images. The visual artifacts present in the output images generated without audio have disappeared the the output when audio modality is introduced. Quantitative metrics also support the hypothesis as average PSNR ($\uparrow 1.65$dB), SSIM ($\uparrow 0.037$) and ERQA ($\uparrow 0.08$)) have increase with the use of audio(refer table 7.1).

We further extend the architecture with axial attention mechanism on the output of CMFS-B. The axial attention supports the proposed architecture in understanding association between different audio frequencies and pixels of each image region. We see further improvement in

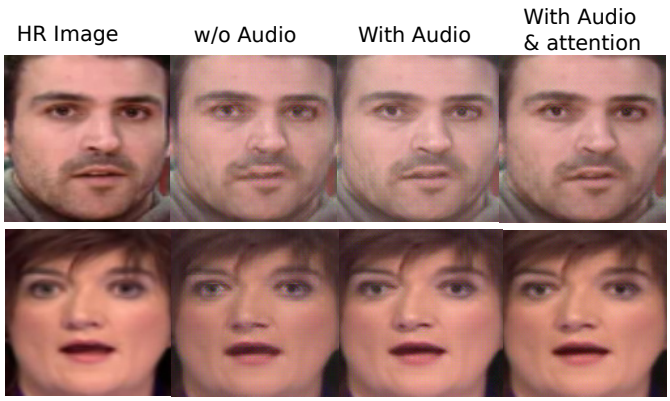


Figure 7.3: Qualitative results on LRW test set showing the importance of audio signal in the proposed network.

Metric	w/o audio	with audio	with audio and attention
PSNR(db)	27.128	28.784	28.960
SSIM	0.862	0.899	0.915
ERQA	0.512	0.592	0.601

Table 7.1: Quantitative results evaluated on LRW test dataset showing the importance of audio signal in the proposed network.

both visual as quantitative results after the use of axial attention. The colors of the facial regions becomes sharper and region boundaries are very distinct.

Loss component analysis: The final function in proposed architecture is combination of five losses (refer 7.4). Initially, we applied the combination of perceptual and adversarial and L1 loss (refer Figure 7.4b and table 7.2a. The results obtained from the network show issue in both color or texture of images. With the introduction of weighted frequency component as the loss function, we can clearly see improvement in texture of generated images (Figure 7.4c and table 7.2b). Still the generated image has small artifacts. As a next step, we introduce lip-reading loss (along with perceptual and adversarial loss) that focuses on generating sharper pixels for mouth region and maintaining temporally consistency (Figure 7.4d and table 7.2c).

The final network uses combination of all five losses to optimize the VFHN. The results in Figure 7.4e and table 7.2d show that there is significant improvement in both the color sharpness and uniformity in the texture of all facial regions.

Different backbone architectures: We studied the effect of different backbones in the section. We used existing SOTA architecture as backbone in place of PHFE-B to demonstrate the significance of proposed architecture. When PHFE-B module in the proposed architecture is replace by generator of SRGAN [69] and EDSR [74], the results show presence of visual artifacts in key facial regions (Figure 7.6b). The quantitative values also show large decline in performance (7.6c).

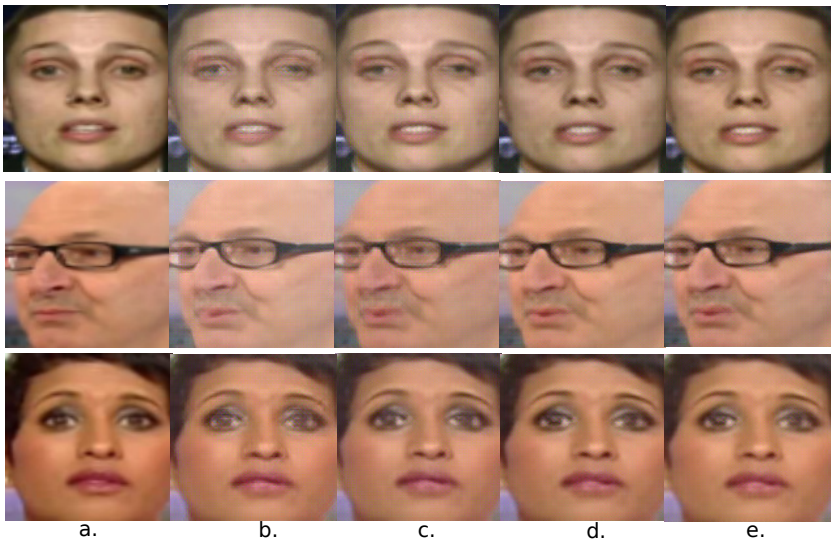


Figure 7.4: Qualitative results for VFHN with different loss functions: a. HR Image, b. perceptual and adversarial loss, c. perceptual, adversarial and weighted frequency loss, d. perceptual, adversarial loss and lip-reading loss, e. perceptual, adversarial loss, weighted frequency loss and lip-reading loss.

Metric	a.	b.	c.	d.
PSNR	26.003	27.892	28.420	28.603
SSIM	0.864	0.893	0.902	0.911
ERQA	0.413	0.471	0.472	0.497

Table 7.2: Average metric values calculated on LRS2 test dataset by using different combinations of loss functions.

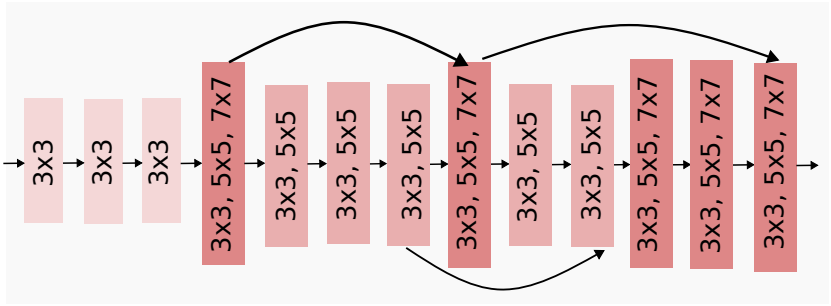


Figure 7.5: mixnet block

After analysing the shortcomings of existing architectures, we propose a generator backbone with multiplex filter structure (refer Figure 7.5). Taking inspiration from mixnet, We design progressive upscaling generator with two mixnet blocks in each stage. Results obtained using this architecture have high resolution with sharp features (refer Figure 7.6d). Yet, there is texture issue in the generated images, since there is color discrepancy between generated and ground-truth image. We resolved the texture issue by using progressive hierarchical feature extractor generator [195]. The empirical metrics shows a marginal improvement from the above discussed generator backbones (refer Figure 7.6e).

7.3.3 Comparison with the SOTA

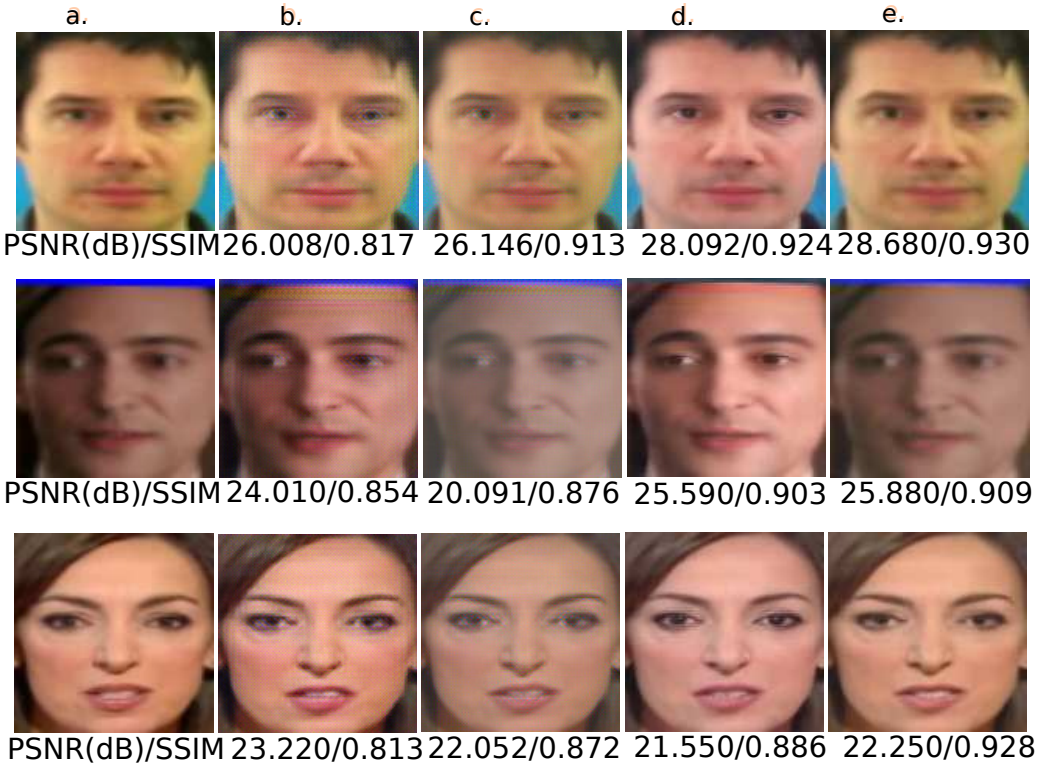


Figure 7.6: Effect of different backbone architectures: a. High resolution image, b. backbone architecture- SRGAN c. backbone architecture- EDSR d. backbone architecture- mixnet and e. image generated with PHFE-B .

Figure 7.7: Qualitative results comparison with $\times 4$ upscaling factor using various datasets: GRID datset (rows: 1), LRW dataset (rows: 2), VFHQ (row: 3 and 4) and Voxceleb (row: 5) (Please zoom in for the better visual comparison).

Method	LRW [202]			LRS2 [193]			GRID [204]			VFHQ [205]			Voxceleb [206]		
	PSNR	SSIM	ERQA	PSNR	SSIM	ERQA	PSNR	SSIM	ERQA	PSNR	SSIM	ERQA	PSNR	SSIM	ERQA
Bicubic	27.734	0.795	0.401	27.991	0.801	0.368	28.192	0.792	0.394	25.942	0.783	0.492	26.103	0.711	0.459
SRGAN [69]	29.262	0.865	0.515	29.631	0.842	0.469	28.715	0.848	0.455	26.753	0.791	0.506	26.147	0.741	0.498
ProgGAN[167]	**29.301**	0.847	0.496	**30.610**	0.831	0.466	29.988	0.820	0.419	26.717	0.808	0.525	**29.595**	0.728	0.511
SICNN [181]	28.031	0.812	0.421	28.172	0.822	0.449	28.042	0.805	0.423	26.872	0.813	0.518	26.911	0.788	0.502
GFPGAN[207]	28.155	0.870	0.490	29.414	0.843	0.485	28.133	0.825	0.424	25.543	0.826	0.534	25.911	0.821	0.573
RealBasicVSR[142]	29.370	0.913	0.583	28.595	0.878	0.484	29.930	0.905	0.504	27.031	0.881	0.581	25.541	0.855	0.598
RealBasicVSR++[208]	29.381	0.905	0.559	28.601	0.863	0.469	30.001	0.909	0.499	**29.416**	0.917	0.449	28.980	0.895	0.402
Ours (mixnet)	28.990	0.911	0.592	28.599	0.893	0.483	30.569	0.913	**0.545**	28.399	0.902	0.572	28.889	0.887	0.550
Ours (main)	28.960	**0.915**	**0.601**	28.603	**0.911**	**0.497**	**30.791**	**0.917**	0.544	29.126	**0.921**	**0.593**	29.003	**0.898**	**0.601**

Table 7.3: Average PSNR (dB), average SSIM and average ERQA numbers comparison on various audio-visual datasets.

For the performance evaluation, we compared proposed VFHN with SOTA super resolution methods: SRGAN [69], ProgGAN [167], SICNN [181], GFPGAN [207], RealBasicVSR [142], RealBasicVSR++ [208] and bicubic interpolation.

Grid speech corpus dataset: The first dataset that we use for comparative analysis is grid speech corpus dataset. We carefully selected 15 videos from the datasets to represent maximum variance in the samples. The SSIM, PSNR and ERQA are used as evaluation and comparison metrics for performance analysis. The empirical results shown in Table 7.3 and visual results shown in row-1 of Figure 7.7 show a clear superiority of the proposed methods over others. Numerically, the proposed model achieved best values for all three metrics. We can also see that the proposed model produces true skin tone and retrieve lost visual features in key areas like mouth. Whereas facial images generated by SRGAN and ProgGAN have artifacts present near the teeth, eyes and hair. The SICNN and RealBasicVSR++ produce blurry images. GPFGAN changes structure of the face and produces appalling eyes area. Although, images generated by RealBasicVSR look sharper than other methods, however after close observation they look highly animated rather than real.

LRS2 dataset & LRW dataset: Similar to above section, we manually selected 15 videos from each datasets making sure that they represent maximum diversity of the present samples. The comparison with SOTA on both the datasets is shown in Table 7.3. Our method achieves highest average SSIM and ERQA numbers for both the datasets. However, PSNR values are slightly lower than ProGAN. If we closely observe the visual results of the ProGAN in row-2 of Figure 7.7, we can clearly see the presence of visual artifacts. Furthermore, according to [69], higher PSNR values don't always employ higher visual quality in images. The output generated using GFPGAN are more inclined toward an image generation rather than the SR task. For example, the model is adding beard to the face which is not present originally.

VFHQ and Voxceleb dataset: The VFHQ and Voxceleb dataset are two high resolution datasets which can be used to qualitatively evaluate the methods due to their better quality images. We again choose 15 videos from each of the datasets for the comparative evaluation. The numerical results of the evaluation are shown in Table 7.3. Here, our proposed methods is able to beat all the recent approaches on SSIM and ERQA metrics. Highest PSNR value is achieved by ProGAN and RealBasicVSR++ on Voxceleb and VFHQ, respectively. Visual results from VFHQ and Voxceleb dastaset are shown in row-3&4 and row-5 of Figure 7.7, respectively. For both the datasets, proposed VFHN shows its ability to retrieve real skin tone and generate the missing region properties like teeth, microphone and glasses. Whereas in methods like RealBasicVSR and GFPGAN, images have overstretched contrast and making image visually unrealistic. The more recent method (RealBasicVSR++) maintains blurriness in the outputs confirming our hypothesis that explicitly learning high frequency component can help in maintaining better region boundaries.

7.4 Conclusion

This work investigates benefits of employing audio signal in video face hallucination task. The chapter empirically verifies that audio signal helps in retrieving the lost visual information and supports in maintaining visual consistency across consecutive frames. We introduce a novel lip-reading loss inspired by visual speech recognition. The loss enables proposed architecture to generate facial images with fine texture details in areas like, mouth, lips, etc. (which is generally absent in the previous SOTA). Further to effectively incorporate salient frequency features, we incorporate a frequency-based loss function in addition to spatial domain loss functions. Our experimental results demonstrate a clear superiority of our proposed model over SOTA.

8 Conclusion and Future Scopes

8.1 Conclusion

Deep neural networks have demonstrated considerable advances in producing high-resolution images. Due to their ability to produce realistic images with precise details and textures, deep learning-based approaches typically GAN-based networks offer an effective solution for super-resolution challenges. By leveraging the adversarial training process, GANs can learn from large datasets and generate high-quality, visually appealing images that are indistinguishable from ground-truth images. Previous GAN-based methods were used to resolve super-resolution problems such as high computational complexity and vanishing gradient. In addition, these methods only rely on 2D information to generate high-resolution images. Therefore, the generated images should include 3D information such as depth and structural information. The frames generated for video super-resolutions have temporal inconsistency.

Hence, in this dissertation, we tried to solve problems present in the previous literature. A detailed description of the proposed frameworks to solve the challenges imposed by the super-resolution problem are as follows:

8.1.1 Semantic Information Based Image Super-Resolution System

Deep learning methods for the super-resolution problem are better than other traditional techniques. However, these methods cannot learn complex spatial structures and high-frequency details, leading to over-smooth results. We developed a novel Generative Adversarial Network based architecture named Residue and Semantic feature-based Dual Subpixel Generative Adversarial Network to solve the super-resolution problem. The generator network is residue and semantic feature-based dual subpixel generative architecture, divided into the premier residual stage and deuxieme residual stage. These two stages are concatenated together to form a two-stage upsampling process, enhancing the model's feature learning capability. Inter and intra-residual connections are made within these two stages, helping to sustain images' high texture details.

Semantic-based information is implanted in a generator to enhance the quality of objects in an image. For embedding semantic information in the generator, feature maps extracted from the pre-trained model are merged with the input image. To stabilize the training process, we introduced spectral normalization in the discriminator. Visual perception and mean opinion score show that the proposed method outperforms the other state-of-the-art methods.

8.1.2 An Efficient Image Super Resolution Model Using Generative Adversarial Networks

This dissertation proposes a novel GAN-based architecture, Super Resolution with Inception Network (SRINet), to solve the high computational complexity, enormous depth, and vanishing gradient problem. The generator architecture of SRINet uses a complex filter structure rather than the linear filter structure to increase the depth and width of the network without increasing the computational cost. Complex filter settings in the architecture help it attain locally distributed information along with hierarchical global information in an image.

Hence, the proposed method approximates the most favorable sparse structures to foster the learning capability of the network. To measure the visual quality of an image, we use a human visual system based visual information fidelity metric. The proposed method outperforms all the state-of-the-art methods qualitatively (perceptually) and quantitatively on other GAN based methods.

8.1.3 Frequency Aware and Semantic Structural Constraint Based Face Hallucination System

In this dissertation, we also address the issue of face hallucination. Most current face hallucination methods rely on two-dimensional facial priors to generate high-resolution face images from low-resolution images. These methods are only capable of assimilating global information into the generated image. Still, there exist some inherent problems in these methods, such as local features, subtle structural details, and depth information is missing in the final output image.

This work proposes a generative adversarial network (GAN) based novel progressive face hallucination (FH) network to address these issues present among current methods. The generator of the proposed model comprises of FH network and two sub-networks, assisting the FH network in generating high-resolution images. The first sub-network leverages explicitly adding high-frequency components into the model. An autoencoder is proposed to generate high-resolution discrete cosine transform (DCT) coefficients to encode the high-frequency components explicitly. The second sub-network is proposed to add three-dimensional parametric information into

the network. This network uses a shape model of 3D morphable models (3DMM) to add structural constraints to the FH network.

8.1.4 Video face hallucination with frequency supervision and cross modality support

Recently, there have been numerous breakthroughs in face hallucination tasks. However, the task remains rather challenging in videos than images due to inherent consistency issues. An extra temporal dimension in video face hallucination makes learning the facial motion throughout the sequence non-trivial. In order to learn these fine spatiotemporal motion details, in this dissertation, we presented a novel cross-modal audio-visual Video Face Hallucination Generative Adversarial Network (VFH-GAN).

The architecture exploits the semantic correlation between the movement of the facial structure and the associated speech signal. Another major issue in current video-based approaches is the presence of blurriness around the key facial regions such as mouth and lips - where spatial displacement is much higher than in other areas. The proposed approach explicitly defines a lip reading loss to learn the fine-grain motion in these facial areas. During training, GANs have the potential to fit frequencies from low to high, which leads to missing the hard-to-synthesize frequencies. Therefore, to add salient frequency features to the network, we add a frequency-based loss function. The visual and the quantitative comparison with state-of-the-art shows a significant improvement in performance and efficacy.

8.2 Future Scope

GANs have been widely used in image super-resolution tasks, and their future scope is promising as they improve accuracy and efficiency. The following are some possible growth areas:

- Generating multiple HR images at different scales from a single LR input allows for more detailed and accurate reconstructions, increasing its usefulness in satellite imagery analysis or medical imaging applications.
- Systems based on generative modeling can be developed to support cross-modality SR. For example, LR infrared (IR) images may be captured in security and surveillance applications due to low light conditions or thermal imaging. However, these LR images may need to provide more details for identification or analysis. These low-resolution IR images can be converted into HR visible-light images using image SR models, providing more detailed information and enhancing the overall image quality.

- Real-time super-resolution is essential in various applications, including video conferences and gaming. GAN-based approaches can be optimized to generate high-quality images in real time by reducing computational complexity and memory requirements, leading to more efficient and practical applications of GAN-based SR.
- GANs can be trained to perform joint super-resolution and image restoration tasks, such as denoising, deblurring, or dehazing, improving the HR images overall quality and making them more usable in downstream applications.

Bibliography

[1] L. Yue, H. Shen, J. Li, Q. Yuan, H. Zhang, and L. Zhang, "Image super-resolution: The techniques, applications, and future," *Signal processing*, vol. 128, pp. 389–408, 2016.

[2] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3365–3387, 2020.

[3] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE signal processing magazine*, vol. 20, no. 3, pp. 21–36, 2003.

[4] S. Sharma, V. S. Bawa, and V. Kumar, "A novel two-stage residual learning based convolutional neural network for image super resolution," *Fundamenta Informaticae*, vol. 168, no. 2-4, pp. 335–351, 2019.

[5] S. Borman and R. L. Stevenson, "Super-resolution from image sequences-a review," in *1998 Midwest symposium on circuits and systems (Cat. No. 98CB36268)*, pp. 374–378, IEEE, 1998.

[6] S. Sayyari, S. Daei, and F. Haddadi, "Blind two-dimensional super-resolution in multiple-input single-output linear systems," *IEEE Signal Processing Letters*, vol. 28, pp. 583–587, 2020.

[7] F. Zhou, W. Yang, and Q. Liao, "Interpolation-based image super-resolution using multisurface fitting," *IEEE Transactions on Image Processing*, vol. 21, no. 7, pp. 3312–3318, 2012.

[8] H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 1, pp. I–I, IEEE, 2004.

[9] S. M. A. Bashir, Y. Wang, M. Khan, and Y. Niu, "A comprehensive review of deep learning-based single image super-resolution," *PeerJ Computer Science*, vol. 7, p. e621, 2021.

[10] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," *arXiv preprint arXiv:1511.08458*, 2015.

[11] V. K. Ha, J. Ren, X. Xu, S. Zhao, G. Xie, and V. M. Vargas, "Deep learning based single image super-resolution: A survey," in *Advances in Brain Inspired Cognitive Systems: 9th International Conference, BICS 2018, Xi'an, China, July 7-8, 2018, Proceedings 9*, pp. 106–119, Springer, 2018.

[12] J. Hatvani, A. Horváth, J. Michetti, A. Basarab, D. Kouamé, and M. Gyöngy, "Deep learning-based super-resolution applied to dental computed tomography," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 120–128, 2018.

[13] A. S. Chaudhari, Z. Fang, F. Kogan, J. Wood, K. J. Stevens, E. K. Gibbons, J. H. Lee, G. E. Gold, and B. A. Hargreaves, "Super-resolution musculoskeletal mri using deep learning," *Magnetic resonance in medicine*, vol. 80, no. 5, pp. 2139–2154, 2018.

[14] H. Temiz and H. S. Bilge, "Super resolution of b-mode ultrasound images with deep learning," *IEEE Access*, vol. 8, pp. 78808–78820, 2020.

[15] J. Jiang, H. Sun, X. Liu, and J. Ma, "Learning spatial-spectral prior for super-resolution of hyperspectral imagery," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1082–1096, 2020.

[16] A. Mehmood, "Deep learning based super resolution of aerial and satellite imagery," in *Automatic Target Recognition XXIX*, vol. 10988, pp. 210–221, SPIE, 2019.

[17] X. Xu, Y. Ma, and W. Sun, "Towards real scene super-resolution with raw images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1723–1731, 2019.

[18] V. Gupta, N. Sambyal, A. Sharma, and P. Kumar, "Restoration of artwork using deep neural networks," *Evolving Systems*, vol. 12, pp. 439–446, 2021.

[19] Z. He, X. Li, and R. Qu, "Video satellite imagery super-resolution via model-based deep neural networks," *Remote Sensing*, vol. 14, no. 3, p. 749, 2022.

[20] Y. Luo, L. Zhou, S. Wang, and Z. Wang, "Video satellite imagery super resolution via convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 12, pp. 2398–2402, 2017.

[21] S. Villena, M. Vega, J. Mateos, D. Rosenberg, F. Murtagh, R. Molina, and A. K. Katsaggelos, "Image super-resolution for outdoor digital forensics. usability and legal aspects," *Computers in industry*, vol. 98, pp. 34–47, 2018.

[22] G. Guarnieri, M. Fontani, F. Guzzi, S. Carrato, and M. Jerian, "Perspective registration and multi-frame super-resolution of license plates in surveillance videos," *Forensic Science International: Digital Investigation*, vol. 36, p. 301087, 2021.

[23] F. Nan, W. Jing, F. Tian, J. Zhang, K.-M. Chao, Z. Hong, and Q. Zheng, "Feature super-

resolution based facial expression recognition for multi-scale low-resolution images," *Knowledge-Based Systems*, vol. 236, p. 107678, 2022.

[24] Z. Qin, W. He, F. Deng, M. Li, and Y. Liu, "Srprid: Pedestrian re-identification based on super-resolution images," *IEEE Access*, vol. 7, pp. 152891–152899, 2019.

[25] Y. Yin, J. Robinson, Y. Zhang, and Y. Fu, "Joint super-resolution and alignment of tiny faces," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 12693–12700, 2020.

[26] P. Rasti, T. Uiboupin, S. Escalera, and G. Anbarjafari, "Convolutional neural network super resolution for face recognition in surveillance monitoring," in *Articulated Motion and Deformable Objects: 9th International Conference, AMDO 2016, Palma de Mallorca, Spain, July 13-15, 2016, Proceedings 9*, pp. 175–184, Springer, 2016.

[27] J. Chen, J. Chen, Z. Wang, C. Liang, and C.-W. Lin, "Identity-aware face super-resolution for low-resolution face recognition," *IEEE Signal Processing Letters*, vol. 27, pp. 645–649, 2020.

[28] R. D. Rakshit, D. R. Kisku, P. Gupta, and J. K. Sing, "Cross-resolution face identification using deep-convolutional neural network," *Multimedia Tools and Applications*, vol. 80, pp. 20733–20758, 2021.

[29] N. S. Ivanov, A. V. Arzhskov, and V. G. Ivanenko, "Combining deep learning and super-resolution algorithms for deep fake detection," in *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, pp. 326–328, IEEE, 2020.

[30] A. Watson, "Deep learning techniques for super-resolution in video games," *arXiv preprint arXiv:2012.09810*, 2020.

[31] K. S. Krishnan and K. S. Krishnan, "Swiftsrgan-rethinking super-resolution for efficient and real-time inference," in *2021 International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, pp. 46–51, IEEE, 2021.

[32] A. Sharma and D. B. Jayagopi, "Towards efficient unconstrained handwriting recognition using dilated temporal convolution network," *Expert Systems with Applications*, vol. 164, p. 114004, 2021.

[33] V. S. Bawa, G. Singh, F. KapingA, I. Skarga-Bandurova, E. Oleari, A. Leporini, C. Landolfo, P. Zhao, X. Xiang, G. Luo, *et al.*, "The saras endoscopic surgeon action detection (esad) dataset: challenges and methods," *arXiv preprint arXiv:2104.03178*, 2021.

[34] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE access*, vol. 7, pp. 19143–19165, 2019.

[35] S. Ray, "A quick review of machine learning algorithms," in *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*, pp. 35–39, IEEE, 2019.

[36] S. Singu, "Comparative analysis of artificial neural networks," *International Journal of Machine Learning for Sustainable Development*, vol. 3, no. 4, 2021.

[37] S. Dong, P. Wang, and K. Abbas, "A survey on deep learning and its applications," *Computer Science Review*, vol. 40, p. 100379, 2021.

[38] S. Qi, X. Ning, G. Yang, L. Zhang, P. Long, W. Cai, and W. Li, "Review of multi-view 3d object recognition methods based on deep learning," *Displays*, vol. 69, p. 102053, 2021.

[39] S. Girisha, U. Verma, M. M. Pai, and R. M. Pai, "Uvid-net: Enhanced semantic segmentation of uav aerial videos by embedding temporal information," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4115–4127, 2021.

[40] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[41] M. Sewak, S. K. Sahay, and H. Rathore, "An overview of deep learning architecture of deep neural networks and autoencoders," *Journal of Computational and Theoretical Nanoscience*, vol. 17, no. 1, pp. 182–188, 2020.

[42] P. Dhruv and S. Naskar, "Image classification using convolutional neural network (cnn) and recurrent neural network (rnn): a review," *Machine Learning and Information Processing: Proceedings of ICMLIP 2019*, pp. 367–381, 2020.

[43] R. Chauhan, K. K. Ghanshala, and R. Joshi, "Convolutional neural network (cnn) for image detection and recognition," in *2018 first international conference on secure cyber computing and communication (ICSCCC)*, pp. 278–282, IEEE, 2018.

[44] C. A. Holt and A. E. Roth, "The nash equilibrium: A perspective," *Proceedings of the National Academy of Sciences*, vol. 101, no. 12, pp. 3999–4002, 2004.

[45] M.-Y. Liu, X. Huang, J. Yu, T.-C. Wang, and A. Mallya, "Generative adversarial networks for image and video synthesis: Algorithms and applications," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 839–862, 2021.

[46] S. Azadi, M. Fisher, V. G. Kim, Z. Wang, E. Shechtman, and T. Darrell, "Multi-content gan for few-shot font style transfer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7564–7573, 2018.

[47] N.-T. Tran, V.-H. Tran, N.-B. Nguyen, T.-K. Nguyen, and N.-M. Cheung, "On data aug-

mentation for gan training," *IEEE Transactions on Image Processing*, vol. 30, pp. 1882–1897, 2021.

[48] X. Xia, X. Pan, N. Li, X. He, L. Ma, X. Zhang, and N. Ding, "Gan-based anomaly detection: A review," *Neurocomputing*, vol. 493, pp. 497–535, 2022.

[49] A. Ramesh, A. S. Rao, S. Moudgalya, and K. Srinivas, "Gan based approach for drug design," in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 825–828, IEEE, 2021.

[50] M. Tao, H. Tang, F. Wu, X.-Y. Jing, B.-K. Bao, and C. Xu, "Df-gan: A simple and effective baseline for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16515–16525, 2022.

[51] C. You, G. Li, Y. Zhang, X. Zhang, H. Shan, M. Li, S. Ju, Z. Zhao, Z. Zhang, W. Cong, *et al.*, "Ct super-resolution gan constrained by the identical, residual, and cycle learning ensemble (gan-circle)," *IEEE transactions on medical imaging*, vol. 39, no. 1, pp. 188–203, 2019.

[52] H. Chen, Q. Xiao, and X. Yin, "Generating music algorithm with deep convolutional generative adversarial networks," in *2019 IEEE 2nd International Conference on Electronics Technology (ICET)*, pp. 576–580, IEEE, 2019.

[53] D. J. Samuel and F. Cuzzolin, "Unsupervised anomaly detection for a smart autonomous robotic assistant surgeon (saras) using a deep residual autoencoder," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7256–7261, 2021.

[54] L. Badino, C. Canevari, L. Fadiga, and G. Metta, "An auto-encoder based approach to unsupervised learning of subword units," in *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 7634–7638, IEEE, 2014.

[55] E. Rituerto-González and C. Peláez-Moreno, "End-to-end recurrent denoising autoencoder embeddings for speaker identification," *Neural Computing and Applications*, vol. 33, no. 21, pp. 14429–14439, 2021.

[56] D. P. Kingma, M. Welling, *et al.*, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, vol. 12, no. 4, pp. 307–392, 2019.

[57] A. Golinski, R. Pourreza, Y. Yang, G. Sautiere, and T. S. Cohen, "Feedback recurrent autoencoder for video compression," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[58] G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, and F. J. Theis, "Single-cell rna-seq denoising using a deep count autoencoder," *Nature communications*, vol. 10, no. 1, p. 390, 2019.

[59] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis*, pp. 4–11, 2014.

[60] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, vol. 184, pp. 232–242, 2016.

[61] Q. Meng, D. Catchpoole, D. Skillicom, and P. J. Kennedy, "Relational autoencoder for feature extraction," in *2017 International joint conference on neural networks (IJCNN)*, pp. 364–371, IEEE, 2017.

[62] Y. Bengio, L. Yao, G. Alain, and P. Vincent, "Generalized denoising auto-encoders as generative models," *Advances in neural information processing systems*, vol. 26, 2013.

[63] G. Zhang, Y. Liu, and X. Jin, "A survey of autoencoder-based recommender systems," *Frontiers of Computer Science*, vol. 14, pp. 430–450, 2020.

[64] V. Shankar and S. Parsana, "An overview and empirical comparison of natural language processing (nlp) models and an introduction to and empirical application of autoencoder models in marketing," *Journal of the Academy of Marketing Science*, vol. 50, no. 6, pp. 1324–1350, 2022.

[65] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, *et al.*, "Recent advances in convolutional neural networks," *Pattern recognition*, vol. 77, pp. 354–377, 2018.

[66] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1646–1654, 2016.

[67] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[68] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient subpixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874–1883, 2016.

[69] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4681–4690, 2017.

[70] W. Shi, J. Caballero, L. Theis, F. Huszar, A. Aitken, C. Ledig, and Z. Wang, "Is the deconvolution layer the same as a convolutional layer?," *arXiv preprint arXiv:1609.07009*, 2016.

[71] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 391–407, Springer, 2016.

[72] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1637–1645, 2016.

[73] Y. Tai, J. Yang, X. Liu, and C. Xu, "Memnet: A persistent memory network for image restoration," in *Proceedings of the IEEE international conference on computer vision*, pp. 4539–4547, 2017.

[74] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 136–144, 2017.

[75] J. Li, F. Fang, K. Mei, and G. Zhang, "Multi-scale residual network for image super-resolution," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 517–532, 2018.

[76] M. Haris, G. Shakhnarovich, and N. Ukita, "Deep back-projection networks for super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1664–1673, 2018.

[77] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback network for image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3867–3876, 2019.

[78] Z. Lin, S. Li, Y. Jiang, J. Wang, and Q. Luo, "Feedback multi-scale residual dense network for image super-resolution," *Signal Processing: Image Communication*, vol. 107, p. 116760, 2022.

[79] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 624–632, 2017.

[80] Y. Wang, F. Perazzi, B. McWilliams, A. Sorkine-Hornung, O. Sorkine-Hornung, and C. Schroers, "A fully progressive approach to single-image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 864–873, 2018.

[81] V. Chudasama, K. Upla, K. Raja, R. Ramachandra, and C. Busch, "Compact and progressive network for enhanced single image super-resolutionâcompresrnet," *The Visual Computer*, vol. 38, no. 11, pp. 3643–3665, 2022.

[82] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," *arXiv preprint arXiv:1511.05440*, 2015.

[83] J. Bruna, P. Sprechmann, and Y. LeCun, "Super-resolution with deep convolutional sufficient statistics," *arXiv preprint arXiv:1511.05666*, 2015.

[84] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

[85] X. Zhu, Z. Li, X. Zhang, H. Li, Z. Xue, and L. Wang, "Generative adversarial image super-resolution through deep dense skip connections," in *Computer Graphics Forum*, vol. 37, pp. 289–300, Wiley Online Library, 2018.

[86] A. Bulat, J. Yang, and G. Tzimiropoulos, "To learn image super-resolution, use a gan to learn how to do image degradation first," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 185–200, 2018.

[87] X. Yu, F. Porikli, B. Fernando, and R. Hartley, "Hallucinating unaligned face images by multiscale transformative discriminative networks," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 500–526, 2020.

[88] D. Song, C. Xu, X. Jia, Y. Chen, C. Xu, and Y. Wang, "Efficient residual dense block search for image super-resolution," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 12007–12014, 2020.

[89] W. Cheng, M. Zhao, Z. Ye, and S. Gu, "Mfagan: A compression framework for memory-efficient on-device super-resolution gan," *arXiv preprint arXiv:2107.12679*, 2021.

[90] J. Liang, H. Zeng, and L. Zhang, "Details or artifacts: A locally discriminative learning approach to realistic image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5657–5666, 2022.

[91] X. Yang, H. Li, X. Li, and T. Li, "Hifgan: a high-frequency information based generative adversarial network for image super-resolution," *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023.

[92] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 286–301, 2018.

[93] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu, "Residual non-local attention networks for image restoration," *arXiv preprint arXiv:1903.10082*, 2019.

[94] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11065–11074, 2019.

[95] J. W. Soh and N. I. Cho, "Lightweight single image super-resolution with multi-scale spatial attention networks," *IEEE Access*, vol. 8, pp. 35383–35391, 2020.

[96] J. Guo, S. Ma, J. Zhang, Q. Zhou, and S. Guo, "Dual-view attention networks for single image super-resolution," in *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2728–2736, 2020.

[97] B. Niu, W. Wen, W. Ren, X. Zhang, L. Yang, S. Wang, K. Zhang, X. Cao, and H. Shen, "Single image super-resolution via a holistic attention network," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pp. 191–207, Springer, 2020.

[98] A. Mehri, P. B. Ardakani, and A. D. Sappa, "Mprnet: Multi-path residual network for lightweight image super resolution," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2704–2713, 2021.

[99] E. Lu and X. Hu, "Image super-resolution via channel attention and spatial attention," *Applied Intelligence*, vol. 52, no. 2, pp. 2260–2268, 2022.

[100] P. Behjati, P. Rodriguez, C. Fernández, I. Hupont, A. Mehri, and J. Gonzàlez, "Single image super-resolution based on directional variance attention network," *Pattern Recognition*, vol. 133, p. 108997, 2023.

[101] J. Zhang, Y. Liao, X. Zhu, H. Wang, and J. Ding, "A deep learning approach in the discrete cosine transform domain to median filtering forensics," *IEEE Signal Processing Letters*, vol. 27, pp. 276–280, 2020.

[102] V. Verma, N. Agarwal, and N. Khanna, "Dct-domain deep convolutional neural networks for multiple jpeg compression classification," *Signal Processing: Image Communication*, vol. 67, pp. 22–33, 2018.

[103] M. M. Islam, V. K. Asari, M. N. Islam, and M. A. Karim, "Single image super-resolution in frequency domain," in *2012 IEEE Southwest Symposium on Image Analysis and Interpretation*, pp. 53–56, IEEE, 2012.

[104] J. Li, S. You, and A. Robles-Kelly, "A frequency domain neural network for fast image super-resolution," in *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2018.

[105] T. Guo, H. S. Mousavi, and V. Monga, "Adaptive transform domain image super-resolution via orthogonally regularized deep networks," *IEEE transactions on image processing*, vol. 28, no. 9, pp. 4685–4700, 2019.

[106] P. Belin, S. Fecteau, and C. Bedard, "Thinking the voice: neural correlates of voice perception," *Trends in cognitive sciences*, vol. 8, no. 3, pp. 129–135, 2004.

[107] D. Hu, F. Nie, and X. Li, "Deep multimodal clustering for unsupervised audiovisual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9248–9257, 2019.

[108] Y. Tian, D. Li, and C. Xu, "Unified multisensory perception: Weakly-supervised audio-visual video parsing," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pp. 436–454, Springer, 2020.

[109] Y. Wu, L. Zhu, Y. Yan, and Y. Yang, "Dual attention matching for audio-visual event localization," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6292–6300, 2019.

[110] H. Xuan, Z. Zhang, S. Chen, J. Yang, and Y. Yan, "Cross-modal attention network for temporal inconsistent audio-visual event localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 279–286, 2020.

[111] Y. Wen, B. Raj, and R. Singh, "Face reconstruction from voice using generative adversarial networks," *Advances in neural information processing systems*, vol. 32, 2019.

[112] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman, M. Rubinstein, and W. Matusik, "Speech2face: Learning the face behind a voice," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7539–7548, 2019.

[113] L. Chen, R. K. Maddox, Z. Duan, and C. Xu, "Hierarchical cross-modal talking face generation with dynamic pixel-wise loss," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7832–7841, 2019.

[114] X. Zhang, X. Wu, X. Zhai, X. Ben, and C. Tu, "Davd-net: Deep audio-aided video decompression of talking heads," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12335–12344, 2020.

[115] G. Meishvili, S. Jenni, and P. Favaro, "Learning to have an ear for face super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1364–1374, 2020.

[116] W. Huang, Y. Chen, L. Mei, and H. You, "Super-resolution reconstruction of face image based on convolution network," in *Advances in Intelligent Systems and Interactive Applications: Proceedings of the 2nd International Conference on Intelligent and Interactive Systems and Applications (IISA2017)*, pp. 288–294, Springer, 2018.

[117] D. Huang and H. Liu, "Face hallucination using convolutional neural network with iterative back projection," in *Biometric Recognition: 11th Chinese Conference, CCBR 2016, Chengdu, China, October 14-16, 2016, Proceedings 11*, pp. 167–175, Springer, 2016.

[118] C. Chen, D. Gong, H. Wang, Z. Li, and K.-Y. K. Wong, "Learning spatial attention for face super-resolution," *IEEE Transactions on Image Processing*, vol. 30, pp. 1219–1231, 2020.

[119] V. Chudasama, K. Nighania, K. Upla, K. Raja, R. Ramachandra, and C. Busch, "E-comsupresnet: Enhanced face super-resolution through compact network," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 2, pp. 166–179, 2021.

[120] H. Huang, R. He, Z. Sun, and T. Tan, "Wavelet-srnet: A wavelet-based cnn for multi-scale face super resolution," in *Proceedings of the IEEE international conference on computer vision*, pp. 1689–1697, 2017.

[121] L. Ying, S. Dinghua, W. Fuping, L. K. Pang, C. T. Kiang, and L. Yi, "Learning wavelet coefficients for face super-resolution," *The Visual Computer*, vol. 37, no. 7, pp. 1613–1622, 2021.

[122] X. Hu, P. Ma, Z. Mai, S. Peng, Z. Yang, and L. Wang, "Face hallucination from low quality images using definition-scalable inference," *Pattern Recognition*, vol. 94, pp. 110–121, 2019.

[123] Z. Feng, J. Lai, X. Xie, D. Yang, and L. Mei, "Face hallucination by deep traversal network," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 3276–3281, IEEE, 2016.

[124] T. Lu, J. Wang, J. Jiang, and Y. Zhang, "Global-local fusion network for face super-resolution," *Neurocomputing*, vol. 387, pp. 309–320, 2020.

[125] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, pp. 318–333, Springer, 2016.

[126] Z. Chen and Y. Tong, "Face super-resolution through wasserstein gans," *arXiv preprint arXiv:1705.02438*, 2017.

[127] B. Huang, W. Chen, X. Wu, C.-L. Lin, and P. N. Suganthan, "High-quality face image generated with conditional boundary equilibrium generative adversarial networks," *Pattern Recognition Letters*, vol. 111, pp. 72–79, 2018.

[128] S. D. Indradi, A. Arifianto, and K. N. Ramadhani, "Face image super-resolution using inception residual network and gan framework," in *2019 7th International Conference on Information and Communication Technology (ICoICT)*, pp. 1–6, IEEE, 2019.

[129] Y. Luo and K. Huang, "Super-resolving tiny faces with face feature vectors," in *2020 10th International Conference on Information Science and Technology (ICIST)*, pp. 145–152, IEEE, 2020.

[130] M. Zhang and Q. Ling, "Supervised pixel-wise gan for face super-resolution," *IEEE Transactions on Multimedia*, vol. 23, pp. 1938–1950, 2020.

[131] Z. Cheng, X. Zhu, and S. Gong, "Characteristic regularisation for super-resolving face

images," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2435–2444, 2020.

[132] S. Ko and B.-R. Dai, "Multi-laplacian gan with edge enhancement for face super resolution," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 3505–3512, IEEE, 2021.

[133] A. Aakerberg, K. Nasrollahi, and T. B. Moeslund, "Real-world super-resolution of face-images from surveillance cameras," *IET Image Processing*, vol. 16, no. 2, pp. 442–452, 2022.

[134] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos, "Video super-resolution with convolutional neural networks," *IEEE transactions on computational imaging*, vol. 2, no. 2, pp. 109–122, 2016.

[135] Y. Jo, S. W. Oh, J. Kang, and S. J. Kim, "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3224–3232, 2018.

[136] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," in *proceedings of the IEEE International Conference on Computer Vision*, pp. 5533–5541, 2017.

[137] S. Y. Kim, J. Lim, T. Na, and M. Kim, "3dsrnet: Video super-resolution using 3d convolutional neural networks," *arXiv preprint arXiv:1812.09079*, 2018.

[138] M. Haris, G. Shakhnarovich, and N. Ukita, "Recurrent back-projection network for video super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3897–3906, 2019.

[139] X. Zhu, Z. Li, X.-Y. Zhang, C. Li, Y. Liu, and Z. Xue, "Residual invertible spatio-temporal network for video super-resolution," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 5981–5988, 2019.

[140] M. Chu, Y. Xie, J. Mayer, L. Leal-Taixé, and N. Thuerey, "Learning temporal coherence via self-supervision for gan-based video generation," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 75–1, 2020.

[141] W. Wen, W. Ren, Y. Shi, Y. Nie, J. Zhang, and X. Cao, "Video super-resolution via a spatio-temporal alignment network," *IEEE Transactions on Image Processing*, vol. 31, pp. 1761–1773, 2022.

[142] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, "Investigating tradeoffs in real-world video super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5962–5971, 2022.

[143] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, "Basicvsr++: Improving video super-resolution with enhanced propagation and alignment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5972–5981, 2022.

[144] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*, pp. 184–199, Springer, 2014.

[145] M. S. Sajjadi, B. Scholkopf, and M. Hirsch, "Enhancenet: Single image super-resolution through automated texture synthesis," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4491–4500, 2017.

[146] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.

[147] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*, pp. 694–711, Springer, 2016.

[148] M. Bevilacqua, A. Roumy, C. Guillemot, and M. L. Alberi-Morel, "Low-complexity single-image super-resolution based on nonnegative neighbor embedding," 2012.

[149] P. Arbelaez, C. Fowlkes, and D. Martin, "The berkeley segmentation dataset and benchmark," *see http://www. eecs. berkeley. edu/Research/Projects/CS/vision/bsds*, 2007.

[150] R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," in *International conference on curves and surfaces*, pp. 711–730, Springer, 2010.

[151] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[152] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[153] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2016.

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

[155] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[156] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2, pp. 416–423, IEEE, 2001.

[157] J.-B. Huang, A. Singh, and N. Ahuja, "Single image super-resolution from transformed self-exemplars," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5197–5206, 2015.

[158] A. Fujimoto, T. Ogawa, K. Yamamoto, Y. Matsui, T. Yamasaki, and K. Aizawa, "Manga109 dataset and creation of metadata," in *Proceedings of the 1st international workshop on comics analysis, processing and understanding*, pp. 1–5, 2016.

[159] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, vol. 7, p. 2, sn, 2005.

[160] C. Han, S. Shan, M. Kan, S. Wu, and X. Chen, "Face recognition with contrastive convolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 118–134, 2018.

[161] K. Nasrollahi and T. B. Moeslund, "Super-resolution: a comprehensive survey," *Machine vision and applications*, vol. 25, no. 6, pp. 1423–1468, 2014.

[162] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1021–1030, 2017.

[163] T. Lu, Y. Guan, Y. Zhang, S. Qu, and Z. Xiong, "Robust and efficient face recognition via low-rank supported extreme learning machine," *Multimedia Tools and Applications*, vol. 77, no. 9, pp. 11219–11240, 2018.

[164] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701–1708, 2014.

[165] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang, "Fsrnet: End-to-end learning face super-resolution with facial priors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2492–2501, 2018.

[166] K. Grm, W. J. Scheirer, and V. Štruc, "Face hallucination using cascaded super-resolution and identity priors," *IEEE Transactions on Image Processing*, vol. 29, pp. 2150–2165, 2019.

[167] D. Kim, M. Kim, G. Kwon, and D.-S. Kim, "Progressive face super-resolution via attention to facial landmark," *arXiv preprint arXiv:1908.08239*, 2019.

[168] X. Yu, B. Fernando, B. Ghanem, F. Porikli, and R. Hartley, "Face super-resolution guided by facial component heatmaps," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 217–233, 2018.

[169] J. Kittler, P. Huber, Z.-H. Feng, G. Hu, and W. Christmas, "3d morphable face models and their applications," in *International Conference on Articulated Motion and Deformable Objects*, pp. 185–206, Springer, 2016.

[170] P. Huber, G. Hu, R. Tena, P. Mortazavian, P. Koppen, W. J. Christmas, M. Ratsch, and J. Kittler, "A multiresolution 3d morphable face model and fitting framework," in *Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016.

[171] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, *et al.*, "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1314–1324, 2019.

[172] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

[173] P. Huber, *Real-time 3D morphable shape model fitting to monocular in-the-wild videos*. University of Surrey (United Kingdom), 2017.

[174] J. R. Tena, M. Hamouz, A. Hilton, and J. Illingworth, "A validated method for dense non-rigid 3d face registration," in *2006 IEEE International Conference on Video and Signal Based Surveillance*, pp. 81–81, IEEE, 2006.

[175] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[176] T. Bolkart and S. Wuhrer, "A robust multilinear model learning framework for 3d faces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4911–4919, 2016.

[177] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

[178] S. Zafeiriou, G. Trigeorgis, G. Chrysos, J. Deng, and J. Shen, "The menpo facial landmark localisation challenge: A step towards the solution," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 170–179, 2017.

[179] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *European conference on computer vision*, pp. 679–692, Springer, 2012.

[180] M. Wang, Z. Chen, Q. J. Wu, and M. Jian, "Improved face super-resolution generative adversarial networks," *Machine Vision and Applications*, vol. 31, pp. 1–12, 2020.

[181] K. Zhang, Z. Zhang, C.-W. Cheng, W. H. Hsu, Y. Qiao, W. Liu, and T. Zhang, "Super-identity convolutional neural network for face hallucination," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 183–198, 2018.

[182] X. Hu, W. Ren, J. LaMaster, X. Cao, X. Li, Z. Li, B. Menze, and W. Liu, "Face super-resolution guided by 3d facial priors," in *European Conference on Computer Vision*, pp. 763–780, Springer, 2020.

[183] D. Zeng, R. Veldhuis, and L. Spreeuwers, "A survey of face recognition techniques under occlusion," *IET biometrics*, vol. 10, no. 6, pp. 581–606, 2021.

[184] D. Yadav and S. Salmani, "Deepfake: A survey on facial forgery technique using generative adversarial network," in *2019 International conference on intelligent computing and control systems (ICCS)*, pp. 852–857, IEEE, 2019.

[185] K. Remya Revi, K. Vidya, and M. Wilscy, "Detection of deepfake images created using generative adversarial networks: A review," in *Second International Conference on Networks and Advances in Computational Technologies*, pp. 25–35, Springer, 2021.

[186] J. Cai, Z. Meng, A. S. Khan, J. OâReilly, Z. Li, S. Han, and Y. Tong, "Identity-free facial expression recognition using conditional generative adversarial network," in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 1344–1348, IEEE, 2021.

[187] J. Lin, Y. Li, and G. Yang, "Fpgan: Face de-identification method with generative adversarial networks for social robots," *Neural Networks*, vol. 133, pp. 132–147, 2021.

[188] J. Wan, J. Li, Z. Lai, B. Du, and L. Zhang, "Robust face alignment by cascaded regression and de-occlusion," *Neural Networks*, vol. 123, pp. 261–272, 2020.

[189] J. He, W. Shi, K. Chen, L. Fu, and C. Dong, "Gcfsr: a generative and controllable face super resolution method without facial and gan priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1889–1898, 2022.

[190] Z. Fan, X. Hu, C. Chen, X. Wang, and S. Peng, "Facial image super-resolution guided by adaptive geometric features," *EURASIP Journal on Wireless Communications and Networking*, vol. 2020, no. 1, pp. 1–15, 2020.

[191] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi, "Real-time video super-resolution with spatio-temporal networks and motion compensation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4778–4787, 2017.

[192] Z.-Q. J. Xu, Y. Zhang, T. Luo, Y. Xiao, and Z. Ma, "Frequency principle: Fourier analysis sheds light on deep neural networks," *arXiv preprint arXiv:1901.06523*, 2019.

[193] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual

speech recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[194] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*, pp. 108–126, Springer, 2020.

[195] S. Sharma, A. Dhall, and V. Kumar, "Frequency aware face hallucination generative adversarial network with semantic structural constraint," *Computer Vision and Image Understanding*, p. 103553, 2022.

[196] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

[197] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6319–6323, IEEE, 2020.

[198] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, 2016.

[199] L. Jiang, B. Dai, W. Wu, and C. C. Loy, "Focal frequency loss for image reconstruction and synthesis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13919–13929, 2021.

[200] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[201] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, "On the spectral bias of neural networks," in *International Conference on Machine Learning*, pp. 5301–5310, PMLR, 2019.

[202] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen, "Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pp. 1–8, IEEE, 2019.

[203] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," tech. rep., CMU-CS-16-118, CMU School of Computer Science, 2016.

[204] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown, "A corpus of audio-visual lombard speech with frontal and profile views," *The Journal of the Acoustical Society of America*, vol. 143, no. 6, pp. EL523–EL529, 2018.

[205] L. Xie, X. Wang, H. Zhang, C. Dong, and Y. Shan, "Vfhq: A high-quality dataset and benchmark for video face super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 657–666, 2022.

[206] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[207] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9168–9178, 2021.

[208] K. C. Chan, S. Zhou, X. Xu, and C. C. Loy, "BasicVSR++: Improving video super-resolution with enhanced propagation and alignment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.