SUPER RESOLUTION (SR) GENERATIVE ADVERSARIAL NETWORKS (GANS) FOR FEATURE EXTRACTION ON DIFFERENT DATASETS

by

Ajay Mishra, M.Sc.

DISSERTATION

Presented to the Swiss School of Business and Management Geneva

In Partial Fulfillment

Of the Requirements

For the Degree

GLOBAL DOCTOR OF BUSINESS ADMINISTRATION

SWISS SCHOOL OF BUSINESS AND MANAGEMENT GENEVA

MARCH 2024

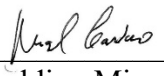
SUPER RESOLUTION (SR) GENERATIVE ADVERSARIAL NETWORKS (GANS) FOR FEATURE EXTRACTION ON DIFFERENT DATASETS

DISSERTATION

by

Ajay Mishra

APPROVED BY



Dr Gualdino Miguel Cardoso, Chair

RECEIVED/APPROVED BY:

Renee Goldstein Osmic

SSBM Representative

DEDICATION

I dedicate this research to my family whose support has made this possible.

ACKNOWLEDGEMENTS

I would like to express my honest gratitude to thesis supervisor Dr. Kamal Malik and Dr. Haitham Nobanee. They always provided supervision and inspiration to keep me on track and on-time. I sincerely thank fellow research students, members of faculty at SSBM and support staff who had made this journey achievable for me. Finally, I'm everlastingly indebted to my family whose support is the very reason I'm able to follow the GDBA program and comprehensively complete this thesis.

ABSTRACT

SUPER RESOLUTION (SR) GENERATIVE ADVERSARIAL NETWORKS (GANS) FOR FEATURE EXTRACTION ON DIFFERENT DATASETS

Ajay Mishra

2024

GLOBAL DOCTOR OF BUSINESS ADMINISTRATION
SWISS SCHOOL OF BUSINESS AND MANAGEMENT

Dissertation Chair: <Chair's Name>
Co-Chair: <If applicable. Co-Chair's Name>

There have been tremendous advancements in using Generative Adversarial Network (GAN) models for computer vision. Despite new advancements in GANs for computer vision, super-resolution (SR) is still considered a challenging research topic in computer vision. Super-resolution has various challenges such as ill-posed inverse problem, complexity growth with increase in up-scaling factor and complexity in assessment of output quality. In recent years, there is a surge of interest in super-resolution methods using Generative Adversarial Networks to solve these challenges. Existing research has been mostly focused on few key techniques and single dataset for feature extraction and calculations. So far, there hasn't been comprehensive research to study various GANs models for existing State-of-the-Art (SOTA) in super-resolution methods for feature extraction on multiple datasets.

The purpose of this research is to study and improve SOTA in super-resolution GANs through extending and modifying various existing architectures and try new architectures for

feature extraction to evaluate results on multiple datasets. This paper focuses on VGG19, VGGFace2 and EfficientNet as pre-trained transfer learning backbones for feature extraction within super-resolution GANs on multiple large challenging face images datasets (CelebA, LFW and Simpson). The result will be evaluated using Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM) metrics.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVATIONS

SR.....…Super Resolution

HR.....…High Resolution

LR.....…Low Resolution

CNN.....…Convolutional Neural Network

GAN.....…Generative Adversarial Network

MSE.....…Mean Squared Error

ESRGAN.....…Enhance Super Resolution Generative Adversarial Network

SRGAN.....…Super Resolution Generative Adversarial Network

RTSRGAN.....…Real Time Super Resolution Generative Adversarial Network

FCSR-GAN.....Face Completion Super Resolution Generative Adversarial Network

RRDB.....…Residual-in-Residual Dense Block

SOTA.....…State of The Art

PSNR.....…Peak Signal to Noise Ratio

SSIM.....…Structural Similarity Index

SISR.....…Single Image Super Resolution

SRCNN.....…Super Resolution Convolutional Neural Network

IRCNN.....…Image Restoration Convolutional Neural Network

ESPCN.....…Efficient Sub-Pixel Convolutional Neural Network

REDNet.....…Residual Encoder Decoder Network

LapSRN.....…Laplacian Pyramid Super Resolution Network

ReLU.....…Rectified Linear Unit

SRDenseNet.....…Super Resolution Dense Net

DDBPN.....…Dense Deep Back Propagation Network

RCAN.....…Residual Channel Attention Network

SRRAM.....…Residual Attention Module for Super Resolution

ZSSR.....…Zero-Shot Super Resolution

SRMD.....…Super Resolution Network for Multiple Degradations

VDSR.....…Very Deep Super Resolution

VGG.....…Visual Geometry Group

MOS.....…Mean Opinion Score

SRPGAN.....…Super Resolution Perceptual Generative Adversarial Network

ProGANSR.....…Progressive Generative Adversarial Network for Super Resolution

CinCGAN.....…Cycle in Cycle Generative Adversarial Network

SAGAN.....…Spatial Attention Generative Adversarial Network

SPSAGAN.....…Self Attention Generative Adversarial Network

RRSB.....…Residual-in-Residual Sparse Block

SPSRGAN.....…Structure Preserving Super Resolution Generative Adversarial Network

LSRGAN.....…Latent Space Regularization Generative Adversarial Network

LCC.....…Lipschitz Continuity Condition

TPGAN.....…Two Pathway Generative Adversarial Network

HLGAN.....…High-to-Low Generative Adversarial Network

AEGAN.....…Auto Embedding Generative Adversarial Network

SDR.....…Standard Dynamic Range

HDR.....…High Dynamic Range

PGAN.....…Progressive Generative Adversarial Network

MR.....…Magnetic Resonance

IPFSRGAN.....ID Preserving Face Super Resolution Generative Adversarial Network

DEM.....…Digital Elevation Model

CHAPTER I:

INTRODUCTION

1.1 Background of the study

Super Resolution (SR) is a process of recovering a high-resolution (HR) image from a low-resolution (LR) image. The new HR image is then referred as Super Resolution (SR) image. SR process has known general problems such as:

- Ill-posed inverse problem – There exist multiple solutions for a given low-resolution image, thus it requires reliable prior information to constrain solution space. A single low-resolution image can correspond to multiple plausible high-resolution versions. The process of down-sampling discards fine-grained details; recovering them from just an LR input is fundamentally ambiguous. SR methods need ways to narrow down the possibilities. This is done by introducing prior knowledge about natural images, this comes through pre-trained models, explicit structural preferences, or the inductive biases of the chosen network architectures.
- Complexity growth with increase in up-scaling factor – With higher up-scaling factors, recovery of missing details become complex and leads to reproduction of wrong data. As the upscaling factor increases (e.g., from 2x to 4x to 8x), the amount of information that needs to be essentially 'invented' by the model grows dramatically. With insufficient guidance or constraints, super-resolution algorithms, especially those relying heavily on deep learning without strong priors, risk producing details that are plausible-looking but entirely inaccurate compared to the true HR original.
- Complexity in assessment of output quality – It is not straightforward to assess quality of output. While pixel-wise metrics like Peak Signal-to-Noise Ratio (PSNR) or Structural Similarity Index (SSIM) provide a starting point, they often fail to correlate with how

humans perceive image quality. A high PSNR image can still lack realistic textures or appear overly blurry. To assess whether an SR output is truly convincing, we need metrics that capture naturalness, fidelity to high-frequency details, and overall aesthetic quality.

Thus, simple up-sampling isn't enough because traditional image resizing (bilinear, bicubic interpolation) essentially averages pixel values. This smooth away fine details instead of recovering them. Interpolation fundamentally can't introduce the information lost during down-sampling. It creates a larger image but not necessarily a sharper one.

To solve these problems efficiently and improve quality of results, Generative Adversarial Network (GAN) have become popular. According to (Goodfellow et al., 2014), GAN is a deep neural network architecture developed based on a game-theory approach where two components of model, Generator and Discriminator compete with each other. Generator create fake images to try to fool Discriminator. Discriminator try not to be fool by learning how to detect fake ones in a better way. This leads to improved learning process for Generator to generate more realistic images.

This competition drives the generator to excel in two ways crucial for super-resolution. First, GANs, due to the discriminator's feedback, encourage the generator to go beyond minimizing pixel-wise discrepancies (as in losses). They inherently prioritize producing images that appear realistic, even if they deviate slightly from the HR ground truth in a pixel-wise sense. Lastly, the "adversarial" focus pushes the generator to become adept at 'filling in' missing high-frequency information in a way that looks natural and aligns with the overall distribution of textures and features it observes in real HR images.

The benefits of GAN approaches in SR are mainly two-fold. First, the core strength of GANs in super-resolution lies in their ability to produce outputs that are perceptually closer to real HR images, with textures and detail that look convincing to the human eye. Lastly, it mitigates Ill-Posed Problem by adversarial training that injects a powerful prior, the generator learns the

characteristics of 'natural-looking' images, reducing the space of plausible solutions to the SR problem.

According to (Anwar et al., 2019; Hu et al., 2019; Cai et al., 2020) Super-resolution GAN models can be categorized in to:

- SRGAN
- ESRGAN
- RTSRGAN
- FCSR-GAN

In these models, the Generator generates super-resolution (SR) image. Discriminator cannot identify it as a genuine high-resolution (HR) image or an artificial SR output. Thus, it leads to generation of HR images with better visible perceptual quality.

The authors of SRGAN (Ledig et al., 2017) proposed to use an adversarial object function to promote SR outputs which is close to the legion of natural images using VGG19 network. SRGAN arguably marks a breakthrough moment. It demonstrated the power of GANs in generating photorealistic, perceptually superior super-resolved images, even when they don't perfectly match the ground truth at a pixel level. They created a multi-task loss function with three parts:

- Mean Squared Error (MSE) loss to encode similarity pixel-wise.
- Metric in terms of distance for perceptual similarity defined over high-level image representation.
- Adversarial loss to balance minimum-maximum game between Generator and Discriminator. It is a standard GAN process objective. (Jessica Li, 2018a)

The authors of VGGFace2 (Cao et al., 2018) introduced a new large-scale face images dataset VGGFace2. The dataset contains more than 3 million images of more than 9 thousand subjects. It has large variations in pose, age, illumination, profession and ethnicity. VGGFace refers

to series of models developed for face recognition by Visual Geometry Group at University of Oxford.

The authors of ESRGAN (Wang et al., 2018a) studied three key components of SRGAN – adversarial loss, perceptual loss and network architecture, and improve each of them to derive an Enhanced SRGAN (ESRGAN). They introduced Residual-in-Residual Dense Block (RRDB) and improved the perceptual loss by using the features before activation, which leads to stronger supervision for brightness consistency and texture recovery. It also Removed batch normalization layers and uses Relativistic GAN for more stable training and potentially sharper output.

The authors of RTSRGAN (Hu et al., 2019) improved on ESRGAN by making it faster at reconstruction speed with acceptable perceptual quality. The aim is to tackle issue of computational efficiency in SRGANs. This has greatly benefitted real time applications such as video feeds. They did it by adding 25 Batch Normalization layers after each convolutional layer. It aims to deliver the perceptual quality closer to ESRGAN-like models while being lightweight enough for real-time applications (e.g., video enhancement on mobile devices).

The authors of FCSR-GAN (Cai et al., 2020) used multi-task learning on face-completion and face super-resolution. They improved face identification performance in case of low-resolution images with occlusion in face images. It uses multi-task learning where Generator performs super-resolution and occlusion removal (inpainting) simultaneously. It uses complex loss function using a combination of adversarial, pixel-wise, perceptual, style, smoothness, and face-specific priors to guide image reconstruction. Tests on public-domain CelebA and Helen databases outperforms SOTA methods in performing face-completion and face super-resolution. It also shows very good conclusive ability in cross- database testing.

The purpose of background study is to understand existing work, datasets used and results measured at PSNR and SSIM metrics. While these metrics have limitations in reflecting perceptual

quality, they provide an initial quantitative measurement of how model and model tuning performs on established datasets. This helps in extending existing GAN models by varying various parameters and testing them on multiple large challenging face images datasets.

1.2 Problem statement

Super Resolution (SR) is a challenging well-known research topic in computer vision despite recent advancements in State-of-the-Art (SOTA) machine learning and deep-learning models. Super-resolution is always in the focus of computer vision community because of use in broad range of applications (Anwar et al., 2019) such as improved reconstructed details of the scenes and objects, object detection in scenes specially mall objects, face recognition, medical imaging, interpretation of images in remote sensing and astronomical images. New super-resolution methods depends up on data-driven deep-learning models to reconstruct details for accurate super-resolution (Anwar et al., 2019).

Super-resolution has various practical challenges (Ledig et al., 2017). These challenges could be tackled using Generative Adversarial Methods (GAN). Generative Adversarial Networks are new breed of models to further advance super-resolution methods and improve results. There has been multitude of research performed on SR using GAN models such as SRGAN (Ledig et al., 2017), FCSR-GAN (Cai et al., 2020), ESRGAN (Wang et al., 2018a) and RTSRGAN (Hu et al., 2019).

Existing research has focused on one or two key techniques and a reference dataset such as Set5-Set14-BSD100 (Wang et al., 2018a), MNIST-TFD (Goodfellow et al., 2014), DIV2K-Flicker2k (Wang et al., 2018a) and OST (Hu et al., 2019) for feature extraction and calculations. This left a gap in research for researchers who are looking for comprehensive research comprising of multiple super resolution GAN models and multiple datasets. This gap prompted me to perform

comprehensive research to study and improve various GANs models for existing State-of-the-Art (SOTA) in super-resolution methods for feature extraction on multiple large challenging face images datasets such as CelebA, LFW and Simpson.

The problem statement is to substitute existing method VGG19 in SRGAN (Ledig et al., 2017) by VGGFace2 (Cao et al., 2018), EfficientNet (Tan and Le, 2019) and to find out if it can improve quantitative metrics (PSNR, SSIM) and observable visual quality of generated images. The research proposal is to use pre-trained transfer learning backbones for feature extraction with multiple large challenging face images datasets such as CelebA (Liu et al., 2018), LFW (Lfw, 2013) and Simpson (alexattia, 2018). The result will be evaluated using PSNR and SSIM metrics.

1.3 Aim and Objectives

The main aim of this research is to study and hopefully improve State-of-the-Art (SOTA) in Super Resolution Generative Adversarial Networks (GANs) for feature extraction on different datasets by performing various experiments. To list the aims of the research:

- Study and improve SOTA in super-resolution GANs through extending and modifying various existing architectures and try new architectures for feature extraction to evaluate results on multiple datasets.
- Focus on VGG19, VGGFace2 and EfficientNet as pre-trained transfer learning backbones.
- Feature extraction within super-resolution GANs on multiple large challenging face images datasets (CelebA, LFW and Simpson).
- Evaluate results using PSNR and SSIM metrics.

The research objectives are formulated based on the aim of this study which are as follows:

- To perform literature review on Generative Adversarial Networks (GAN) models for Super Resolution (SR).

- To perform comparative analysis using transfer learning for feature extraction.
- To propose GAN based SR method.
- To implement the proposed technique on multiple datasets and compare with State-of-the-Art (SOTA) work.

1.4 Scope of the study

The scope of this research is:

- Literature review on existing SR methods with strong focus on Generative Adversarial Networks (GAN) models.
- Successful use of VGG19, VGGFace2 and EfficientNet as pre-trained transfer learning backbones.
- Development of new GAN based super-resolution architecture by changing feature extractor and changing various hyper parameters.
- Evaluate model performance using Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM) metrics on various large challenging face image datasets.

1.5 Significance of the study

By performing this study on various GANs models for existing State-of-the-Art (SOTA) in super-resolution methods and hopefully improving feature extraction by using pre-trained transfer learning backbones on multiple large challenging face images datasets, this study would help fellow researchers to further improve GAN models for SR in the field of feature extraction, improve perceptual visual quality of generated image and higher upscaling factors.

1.6 Structure of the study

This study starts with introduction of Super Resolution, Generative Adversarial Network and Feature Extraction space of GAN based SR methods. After that, it shows overview of SR methods, overview of deep convolutional neural network work for SR methods, review of recent GAN based SR methods, review of real-world applications using GAN based SR methods and present PSNR and SSIM metrics values used in CNN and GAN models.

It also lists gaps identified in existing research of GAN based SR methods. Later, it discusses technical implementation of proposed model, various loss functions and upscaling factors, dataset description and model evaluation parameters. At end, it lists all references used in this study. Figure 1.1 summarizes structure of thesis in concisely and neatly.

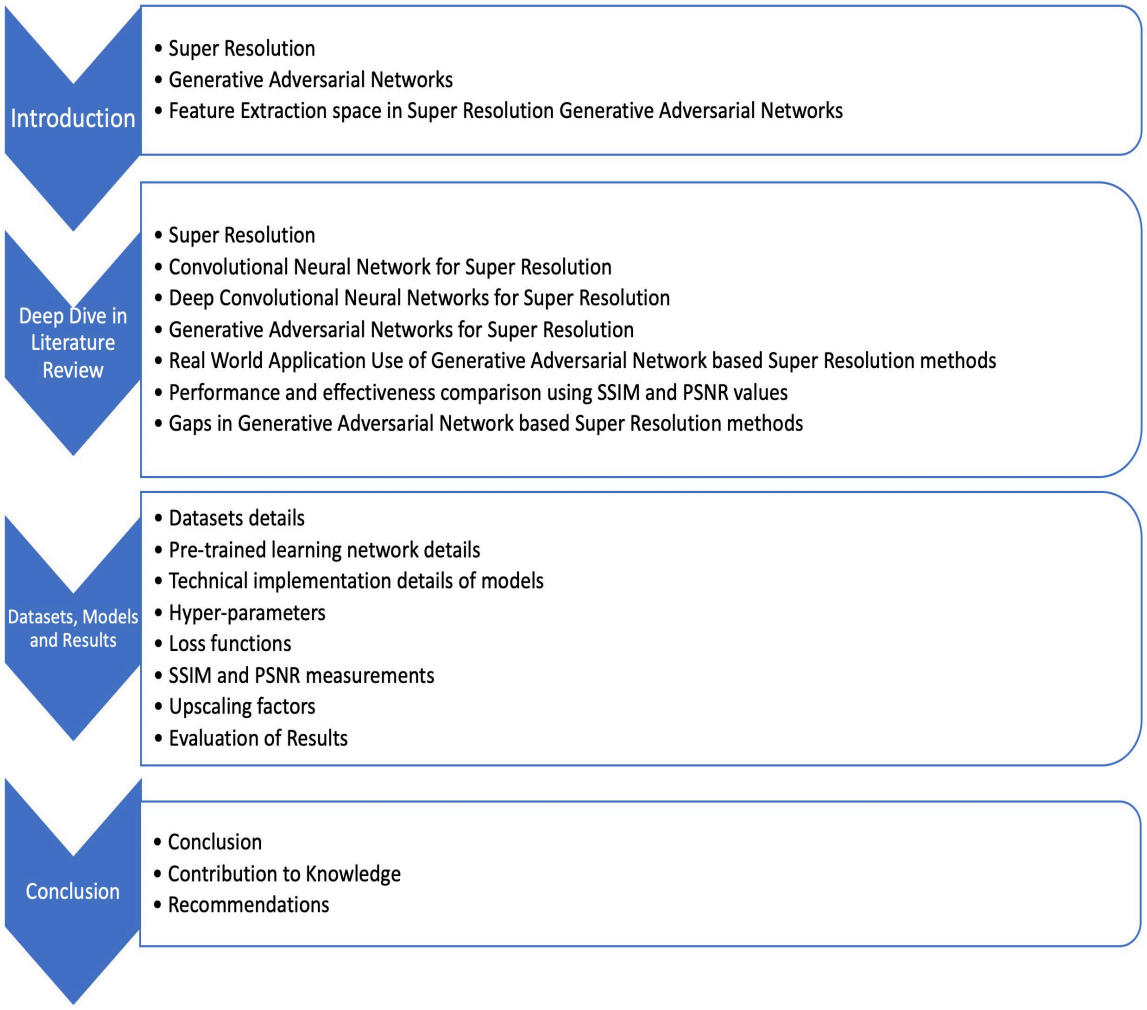


Figure 1.1
Structure of the thesis

CHAPTER II:

LITERATURE REVIEW

2.1 Introduction

Super Resolution (SR) is a way to increase visual quality and details of a low-resolution image by converting it to a high-resolution image. Super Resolution is also known by other names such as image interpolation, image up-sampling, image enlargement and image scaling (Anwar et al., 2019). SR images provides enhanced details of the scenes and objects which are really useful in many new generation electronic devices. An efficient process of SR can drastically reduce requirements on computing power, network bandwidth and storage. SR has many applications in real world scenario such as object detection in scenes, medical imaging, interpretation of images in remote sensing, forensic science, face recognition in surveillance videos, astronomical images and face attribute editing based mobile apps.

The process of generating a new high-resolution image can be completed by using a single low-resolution image or multiple low-resolution images. In this paper, I would mainly focus on Single Image Super Resolution (SISR). Single Image Super Resolution is still a very challenging problem in computer vision when up-scaling requirement are higher than 4x. SR is also a challenging task for face images, face recognition, face parsing and face attribute learning where the requirement is to obtain high-resolution non-occluded images from low-resolution images having occlusions (Cai et al., 2020).

Initial SR methods were developed using traditional classical ways. With the advent of deep learning methods, classical methods are outperformed by a big margin with deep learning-based methods. According to (Anwar et al., 2019), SR methods are of five types: prediction methods, statistical methods, edge based methods, deep learning methods and patch based methods. In this

literature review, I would mainly focus on methods which use deep neural networks, especially Generative Adversarial Network (GAN) models for SISR.

GAN is a framework (Goodfellow et al., 2014) of estimating generative models using an adversarial method. In this process, training for two models run simultaneously. The first model is a generative model G for capturing data distribution and second model is discriminative model D for estimating the probability of underlying sample. GAN framework basically works on a game theory of minimax two player game.

The task of generative models G is to maximize the probability of discriminative model D making a mistake. The generative model G is facing against an adversary discriminative model D. The task of discriminative model D is to learn to determine if a sample is from data distribution or a sample is from model distribution. This competition between generative model G and discriminative model D in this game drives both models to improve in generating realistic images and identifying fake images, respectively. Both models are multilayer perceptron. The aim is to train both models with only greatly successful dropout algorithms and backpropagation. The aim is to sample from generative model using only forward propagation.

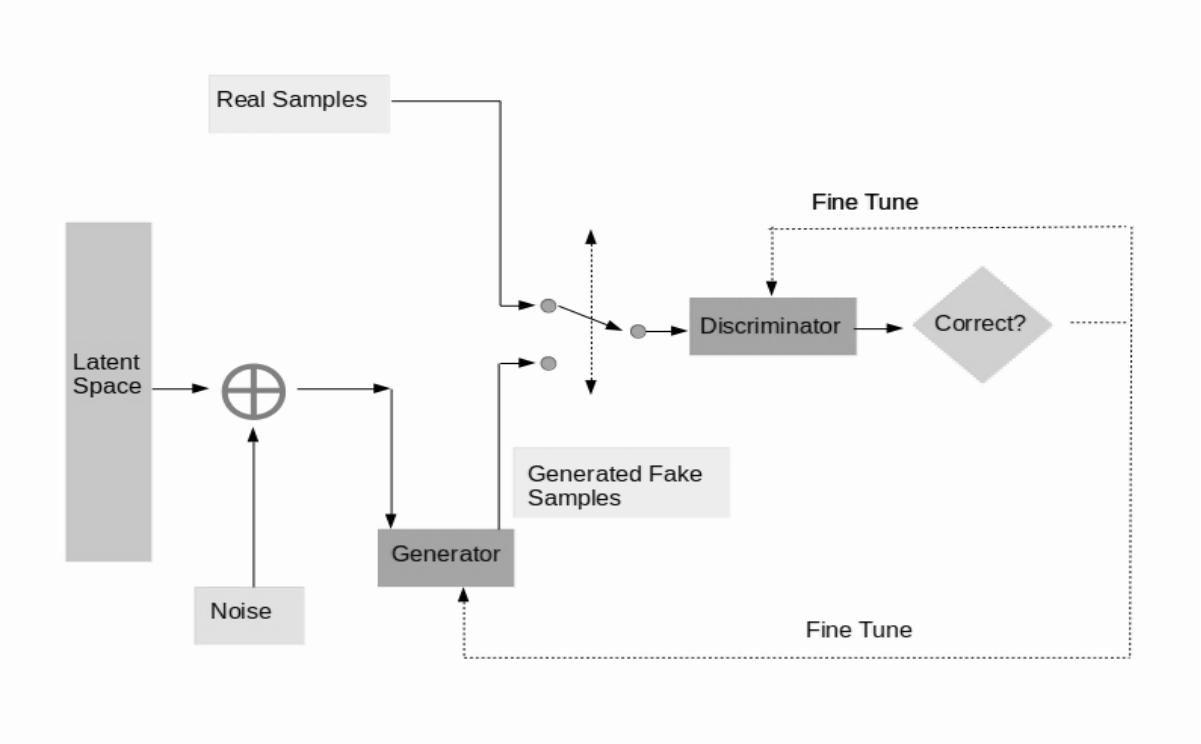


Figure 2.1
GAN conceptual architecture.

GAN have various computational advantages such as it doesn't require inference during learning, it never requires Markov chains, it uses backdrops to obtain gradients and it allows to add wide variety of functions in to model. GAN statistical advantage is that generator network updates only with gradients passing through discriminator and it can represent very sharp degenerate distributions evenly. These, along with ability of GAN to generate new content makes it a very good candidate for SR work in real life applications.

Feature extraction is a sampling problem where perfect reconstruction of desired features are wanted. It is one of the most important aspects of SR models, especially useful in cases of image processing, medical image analysis, face attributes, face identification, face mapping, face verification, face alignment, face normalization, face synthesis, tone mapping and small object identification. Feature Extraction is mostly embedded in convolutional operations layers. The

quality of output images largely depends on the accuracy of extracted features. The generative aspects of GAN based models deal with generation of new images using structural feature information present in training data and discriminative aspect of GAN models learns structural feature information from the original images to rule out anomalies present in the abnormal generated image.

Finally summing up everything so far; SR is a critically acclaimed computer vision and image processing task. Various methods have been proposed to tackle SR challenges and GAN based models are front runner. GAN based models have worked wonderfully on large complex datasets and in practical real-life use cases. They offer plenty of opportunity to tweak and introduce new functions.

This literature review focuses on these points:

- An overview of SR methods.
- An overview of deep convolutional neural network work for SR.
- A review of recent GAN based SR methods.
- A review of real-world applications using GAN based SR methods.
- Tabular representation of PSNR and SSIM metrics values used in CNN and GAN models.
- Discussion and challenges.

2.2 An overview of Super Resolution methods

In this section, the aim is to present an overview of various deep learning-based SR methods.

2.2.1 Linear network methods

Earliest deep learning-based SR methods used simple network designs and are called linear network designs. According to (Anwar et al., 2019), linear networks have only a single path for signal flow and doesn't have any skip connections or multiple-branches. The input flows sequentially from initial layers to other layers comprised of several convolutional layers stacked on top of each other. In this section, I will discuss Super Resolution Convolutional Neural Network (SRCNN), Image Restoration Convolutional Neural Network (IRCNN) and Efficient Sub-Pixel Convolutional Neural Network (ESPCN).

- SRCNN: It is an early up-sampling design. It is the first successful and pioneering work in using just convolutional layers to perform SR. It first performs up-sample on LR input image to match with desired HR output image size. It uses computationally intensive Bicubic Interpolation for up-sampling operation. Each convolutional layer (total three) is followed by rectified linear unit (ReLU). First layer is a feature extraction layer and it creates feature map of input image. SRCNN (Dong et al., 2016) uses Mean Squared Error (MSE) loss function to minimize difference between ground truth high-resolution image and output reconstructed high-resolution image.

  SRCNN's core structure can be understood as a three-step process. Patch extraction and representation layer extracts overlapping patches from the bicubically upscaled LR image and represents them as feature vectors. Patch size and stride are key hyperparameters here, influencing the level of detail captured for reconstruction. The middle convolutional layer performs a non-linear transformation on the extracted patches. This is where the

network learns a more complex mapping from the lower-dimensional representation to a higher-dimensional one, aiming to recover detail lost during the initial up-sampling. The final convolutional layer aggregates the processed patches to produce the final HR image. The goal is to combine the information from the non-linear mappings to fill in missing high-frequency details and generate a visually plausible output.

SRCNN relies on bicubic interpolation as a pre-processing step to upscale the LR image. While efficient, bicubic interpolation introduces smoothing and potential artifacts into the upscaled image. More recent SRGAN approaches often integrate up-sampling directly into the network architecture. SRGANs frequently compare against bicubic interpolation results to demonstrate the power of learned up-sampling.

The use of ReLU activations was a common practice in early deep learning models. ReLU helps prevent gradients from becoming too small during backpropagation, improving training stability in deeper networks compared to earlier activation functions like sigmoid or tanh. The simple piecewise linear form makes it computationally inexpensive to evaluate.

MSE is a standard loss function used to compare the pixel-wise difference between the reconstructed HR image and the ground truth. However. MSE-optimized images may be blurry even if they have low pixel-level error, as it does not directly correspond to how humans perceive image quality. SRGANs often introduce more sophisticated loss functions, incorporating adversarial losses and perceptual metrics to drive the generation of visually sharper and more realistic outputs.

SRCNN set the stage for more complex super-resolution models. The general idea of using CNNs for end-to-end super-resolution, introduced by SRCNN, paved the way for the development of SRGAN architectures. While SRCNN directly maps LR patches to HR equivalents, subsequent models demonstrated the advantages of residual learning, where

the network learns to predict the difference between LR and HR, improving efficiency. Lastly, newer approaches have explored trainable up-sampling layers, going beyond SRCNN's reliance on bicubic interpolation.

- IRCNN: This method also uses early up-sampling design, similar to SRCNN. While IRCNN shares the early up-sampling strategy with SRCNN, it makes several advancements focused on enhancing denoising capabilities and computational efficiency during the super-resolution process. It proposed a set of denoisers to be used jointly. IRCNN explicitly frames super-resolution as an image restoration task where the goal is not just to magnify but also to remove noise and artifacts that are amplified during up-sampling.

  This emphasis on denoising influences several architectural choices. Instead of a single network pipeline, IRCNN trains a set of 25 CNN-based denoisers, each specialized in handling a specific noise level. This ensemble allows the model to dynamically adapt to varying levels of noise that might be present in different LR images. Combining the predictions of multiple denoisers helps reduce the risk of overfitting to any single noise pattern and could improve the overall quality of the reconstructed image.

  The CNN based denoiser is composed of a stack of 7 convolutional layers interleaved with ReLU non-linearity layers and batch normalization. Deeper networks have the potential to learn more complex denoising mappings. Batch normalization helps reduce internal covariate shift within network layers, promoting faster and more stable learning.

  It also uses a technique called Half Quadratic Splitting (HQS) to uncouple regularization and fidelity terms in observation model. HQS breaks down the image restoration optimization problem into sub-problems that can be solved separately, one focused on fidelity to the original data and another on regularization (imposing smoothness or other priors. HQS allows for an iterative approach where these sub-problems are solved

repeatedly, refining the solution at each step. In some cases, HQS can lead to faster convergence and more flexible optimization strategies compared to directly solving the full restoration problem. Denoisers (a set of 25) are trained with various noise level to collectively used for image restoration task has strong performance on image SR.

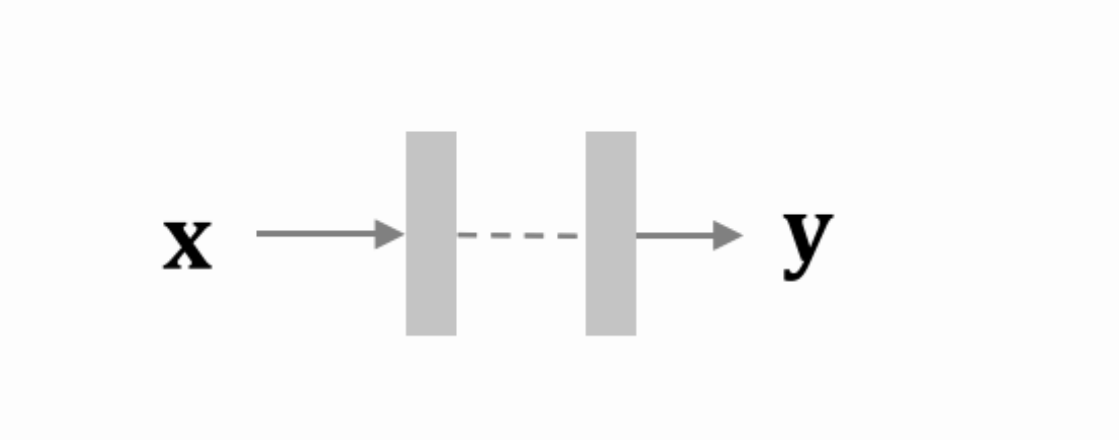


Figure 2.2
Simple representation of network architecture for SRCNN and IRCNN (Anwar et al., 2019).

Training an ensemble of denoisers and the iterative nature of HQS might increase computational overhead compared to a simpler single-network model like SRCNN. While IRCNN marked progress, subsequent SRGANs have largely shifted towards perceptual loss functions and adversarial training which better capture human notions of image quality.

ESPCN: It is a late up-sampling design. Late up-sampling is an efficient approach with low memory footprint compared to computationally expensive early up-sampling design of SRCNN and IRCNN. It is a fast SR method and it can operate in real time for images and videos. Core design philosophy focuses on computational efficiency and reduced memory footprint, making it a landmark in real-time super-resolution.

It performs feature extraction in LR space and then uses a sub-pixel convolutional layer at end to combine LR feature maps while simultaneously doing projections to high

dimensional space to reconstruct high-resolution image. By performing feature extraction and the majority of computations in the low-dimensional LR space, ESPCN avoids processing redundant information that would be introduced by early up-sampling. This late up-sampling strategy directly translates to fewer operations and, consequently, a faster model. Unlike fixed interpolation techniques, the sub-pixel convolution layer has trainable parameters. This allows the network to learn an optimal up-sampling mapping directly from the data, potentially leading to superior results.

It uses L1 loss to train network. SRCNN relies on Mean Squared Error (MSE) loss, ESPCN adopts the L1 loss. L1 loss is less sensitive to outliers (large pixel-wise errors) compared to MSE. This shows improve training stability for dataset containing noisy examples.

It has strong SR performance for images and videos. It has significantly reduced memory and compute because of feature extraction in LR space. The combined effect of late up-sampling and the efficient sub-pixel convolution layer makes it remarkably fast. ESPCN opened up potential for super-resolving videos in real-time, as opposed to frame-by-frame processing with slower models.

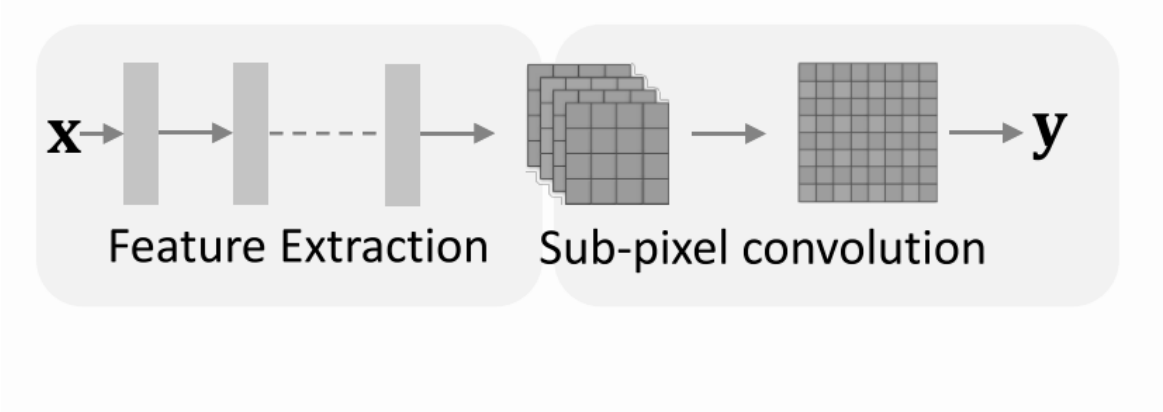


Figure 2.3
Simple representation of network architecture for ESPCN (Anwar et al., 2019).

2.2.2 Residual network methods

The introduction of residual networks was a significant breakthrough in deep learning, enabling the development of substantially deeper architectures without sacrificing performance. It addressed core problems such as vanishing gradients; in very deep traditional neural networks, gradients flowing during backpropagation can become increasingly small in the earlier layers. This makes it difficult to update the weights effectively, hindering training convergence. Residual networks introduce skip connections that bypass one or more network layers. This provides an alternative path for gradients to flow, directly connecting earlier layers to later ones. The network then learns a residual mapping, the difference between the desired output and the input.

The residual networks method uses skip connections in network design. This helps to avoid gradients vanishing and provide a way to create very deep networks (Anwar et al., 2019). In this method, the task is to learn residue (high frequency) between ground truth and input (He et al., 2016). Skip connections help to ensure that even very deep networks remain trainable. In the worst-case scenario, if a residual block fails to improve upon the input, it can simply default to an identity mapping, preventing performance degradation.

There are two types of residual networks, single state (composed of single network) and multi stage (composed of multiple subnets). In this section, I will discuss a multi stage residual network called Residual Encoder Decoder Network (REDNet).

REDNet: It exemplifies how residual principles can be integrated into a super-resolution architecture with a specific emphasis on image reconstruction quality. It is mainly composed of convolutional and symmetric deconvolutional layers with addition of ReLU after each of these layers. The convolutional and deconvolutional layers can be seen as an encoder-decoder structure. The convolutional layers progressively down-sample the input image while extracting increasingly

abstract feature representations. Deconvolutional layers aim to reconstruct the HR image by up-sampling these feature representations.

It has skip connections between convolutional and symmetric deconvolutional layers. Skip connections enable the network to directly propagate high-frequency details from the earlier layers to the output image, helping to maintain sharpness and reduce blur. These connections improve gradient flow throughout the deep network, promoting proper convergence during training. The input to the network is Bicubic Interpolation images and output from final deconvolutional layer is a high-resolution image.

It sums up feature maps of convolutional layer with the output of mirrored deconvolutional layer and then applies non-linear rectification. This serves to combine detailed low-level features from the encoder with the up-sampled features from the decoder, potentially enriching the image reconstruction process. To achieve convergence, it minimizes L2 norm between ground truth and output of the system. The best performing architecture of REDNet used 30 weight layers (each with 64 feature maps) and used Mean Square Error as loss function.

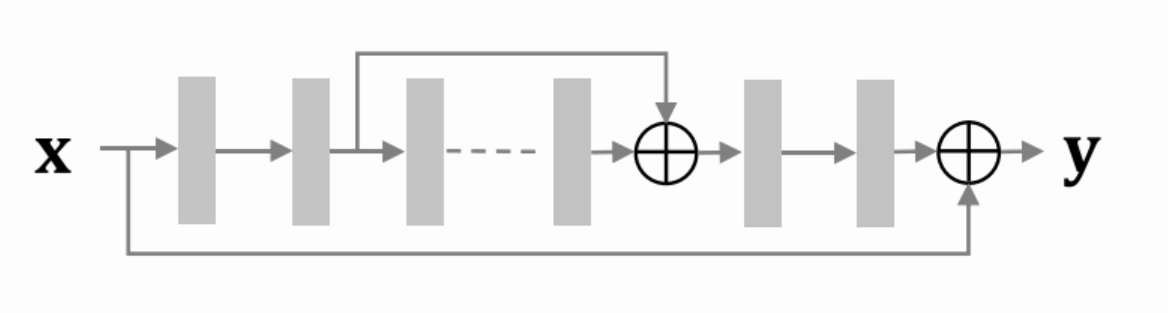


Figure2.4
Simple representation of network architecture for REDNet (Anwar et al., 2019).

2.2.3 Progressive reconstruction methods

Generally, convolutional neural network methods predict output in one step. Thus, it makes them not feasible for large scaling factors (8x). To deal with large up-scaling of image resolution, progressive reconstruction methods are invented. They predict output in multiple steps. In this section, I will discuss Deep Laplacian Pyramid Super Resolution Network (LapSRN).

- LapSRN: It addresses this challenge with a multi-stage, progressive framework inspired by the concept of Laplacian pyramids in classic image processing. It uses a pyramidal framework consisting of three subnets to progressively predict residual images up to factor of 8x. A Laplacian pyramid decomposes an image into several levels, where each level represents the residual differences compared to a down-sampled version. These residuals contain the high-frequency detail lost during down-sampling.

  The first subnet output is a residue of 2x, second subnet output is a residue of 4x and last subnet output is a residue of 8x. These residual images are added to related up-sampled scaled images to get final super-resolved images. Breaking down the challenging 8x super-resolution problem into smaller, more manageable steps helps the network focus on learning the necessary mappings incrementally. Predicting residuals instead of the full HR image in one shot potentially reduces the complexity of the task for each subnet. As each subnet produces an output, loss functions can be applied at multiple stages, providing stronger learning signals to guide optimization. Residual prediction branch is termed as feature extraction branch.

  The network consists of three types of elements. First is convolutional layer, followed by leaky ReLU and finally deconvolutional layer. It uses Charbonnier loss function (differentiable variant of L1 loss function). The loss is put at every subnet, thus creating a multi-loss structure. This loss function is a differentiable variant of L1 loss. Its

robustness to outliers can be beneficial during SR training. Out of three distinct models provided by LapSRN, single MS-LapSRN performance is best. This points to the potential advantages of a consistent progressive approach over variations using multi-scale feature extraction branches. LapSRN provides great speed and accuracy for SR up to 8x.

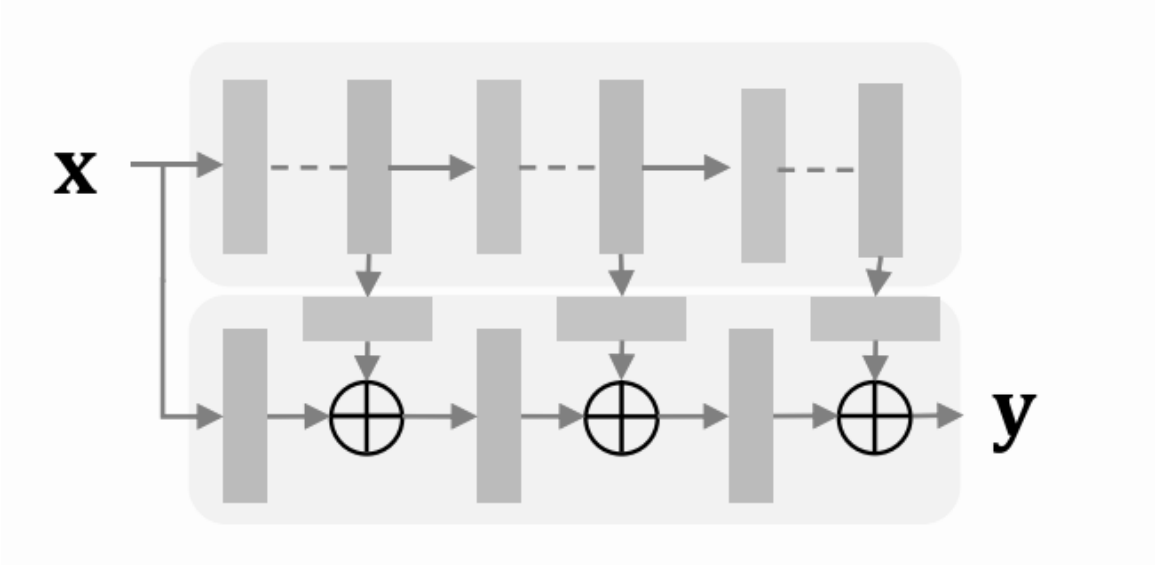


Figure2.5
Simple representation of network architecture for LapSRN (Anwar et al., 2019).

2.2.4 Densely connected methods

To further improve performance of SR methods, densely connected convolutional network layers are proposed. The main idea in densely connected design is to combine available hierarchical info with network depth to achieve richer feature space and high flexibility. In this section, I will discuss Super Resolution Dense Net (SRDenseNet) and Dense Deep Back Propagation Network (DDBPN).

- SRDenseNet: This method uses dense connections as a layer operating on the output from all previous layers. Direct connections ensure that feature maps from early layers are readily accessible to later layers, boosting the network's capacity to propagate and refine relevant

information for the SR reconstruction task. Secondly, by combining features from different levels of abstraction, the network potentially learns a richer and more discriminative representation, which proves beneficial for recovering complex textures and details within super-resolved images. This flow of information from low-level to high-level feature layer speeds up training process, enables learning compact models and avoids vanishing gradient problem. Dense connections provide numerous alternative routes for gradients to flow during backpropagation. It provides implicit "Deep Supervision", since each layer receives information from the preceding layers, in a sense, intermediate layers receive some degree of supervision signal from the later layers. This can help with convergence.

At the rear part of network, it upscales input by using deconvolutional layers. SRDenseNet uses Mean Square Error L2 loss to train the model. Out of three variants proposed, the variants with all feature maps gives best performance because complementary features are encoded at multiple layers in network. This supports the hypothesis that dense connections facilitate the accumulation of complementary and diverse features, leading to better HR reconstructions.

SRDenseNet shows a steady improvement in performance compared to models not using dense connection between layers.

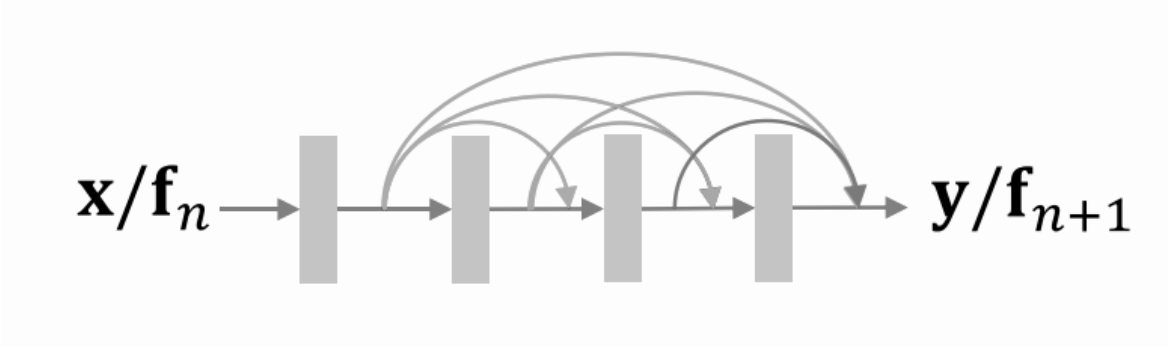


Figure2.6
Simple representation of network architecture for SRDenseNet (Anwar et al., 2019).

- DDBPN: It takes inspiration from conventional SR methods to iteratively perform back-projections. In classic SR, an initial up-sampling is progressively refined by comparing it to the original LR image and using the error to guide further improvement. It aims to integrate a similar iterative feedback mechanism within a deep learning architecture. This helps it learn feedback error signal between low-resolution and high-resolution images.

  The feedback mechanism vastly helps in achieving better SR result. To achieve this purpose, the network consists of densely connected series of up sampling and down sampling layers to combine high-resolution images from multiple depth in network to produce final SR output. Up-sampling blocks increase the spatial resolution of feature maps. Down-sampling blocks reduce resolution but increase the number of feature channels. This combination of up- and down-sampling forces the network to learn representations at multiple scales and levels of detail.

  Another important feature of this architecture is residual signal. It computes a residual between the up-sampled feature maps and a correspondingly upscaled version of the input. This residual carries the error signal, informing the network how to refine its reconstruction iteratively. The residual signal in the up-sampled feature map gives error feedback and pushes network to learn finer details. This helps in producing great looking SR output. DDBPN uses standard L1 loss function to train network.

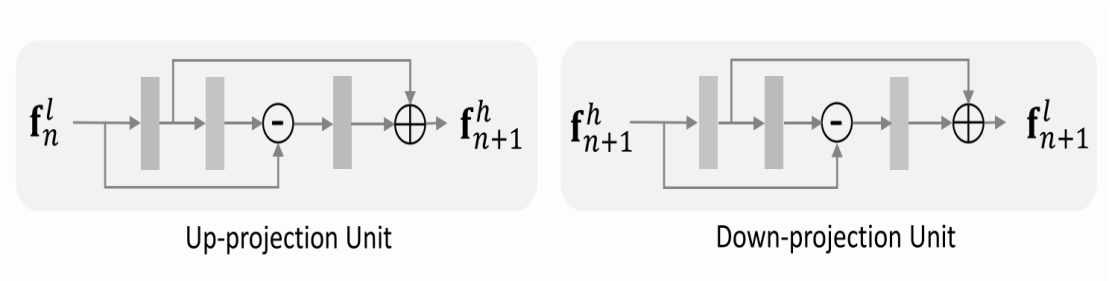


Figure2.7
Simple representation of network architecture for DDBPN (Anwar et al., 2019).

2.2.5 Attention based methods

We have discussed linear network methods, residual network methods, progressive reconstruction methods and densely connected methods. All of these network designs give uniform importance to all spatial locations and channels for SR. However, in several cases, selection-based approach works better by selectively attending to few features at any given network layer. This gives birth to Attention based models. It runs on idea that not all features are essential for SR. Different features have varying importance. Attention in deep learning involves introducing mechanisms that enable the network to dynamically 'focus' on the most relevant parts of its input or intermediate feature maps while downplaying others. This selective emphasis can be across spatial dimensions, feature channels, or both. Attention allows the network to prioritize regions containing high-frequency details relevant for SR. Similarly, among the numerous feature channels processed by convolutional layers, some might be more critical in recovering specific textural patterns or edges. Channel-wise attention can help the network highlight these essential features. Lastly, by focusing on the most relevant regions and features instead of processing everything uniformly, attention mechanisms have the potential to improve computational efficiency.

By coupling with deep neural networks, recent attention-based models have shown great SR performance. Attention mechanisms can provide a degree of interpretability into what the SR

network deems important, potentially revealing which image regions or features are crucial for the reconstruction process. Attention-equipped SR models could be more adaptable to different image domains or target upscaling factors by learning to shift focus dynamically. In this section, I will discuss Residual Channel Attention Network (RCAN) and Residual Attention Module for Super Resolution (SRRAM).

- RCAN: It is deep convolutional neural network (CNN) for SR. The main component of network architecture consists of a local residual block acting as a selective attention over channel maps with a channel attention mechanism to filter collapsed activation and a recursive residual design with residual connections within every block of global residual network. The network is composed of several residual groups, each containing internal residual blocks with connections among them. A long skip connection directly links the beginning and end of the Residual-in-Residual structure.

  It allows multiple pathways from initial to final layers for information flow and it allows network to focus on selective feature maps to not only model relationship between feature maps, but also allows network to focus on feature maps which are important for the end task. This helps in combating vanishing gradients by making training very deep architectures more feasible. Lastly, combining information from early and later layers provides a richer representation for the super-resolution task.

  RCAN uses L1 loss function to train network. RCAN leads to better performance in SR compared to past approaches such as IRCNN. RCAN achieved significant performance gains over earlier models, demonstrating the power of its deep, attention-equipped design. RCAN is very effective for low-level vision tasks but has relatively high computational complexity compared to LapSRN. The complexity of RCAN, due to its depth and attention mechanism, translates to higher computational requirements. The trade-

off between performance and efficiency is a vital consideration in real-world applications. RCAN's success spurred a wave of research into sophisticated attention mechanisms for SR, exploring spatial attention, self-attention, and non-local attention.

- SRRAM: In this model, a residual attention module is proposed. It introduces a specialized residual attention module designed to refine feature representations for superior super-resolution results, demonstrating the continued value of attention-based mechanisms within the SR domain.

SRRAM consist of three parts. First is feature extraction, the initial stage of SRRAM extracts basic features from the input LR image using convolutional layers. Second is feature upscaling, this core stage contains a series of Residual Attention Modules responsible for learning refined, high-dimensional feature representations and progressively upscaling the feature maps. Finally, third feature is reconstruction, It aggregates the information from the upscaled features to reconstruct the final HR image.

The feature upscaling part consists of residual attention modules. The residual attention module is formed of residual blocks, spatial attention and channel attention to learn dependencies of inter-channel and intra-channel. SRRAM integrates both spatial and channel attention mechanisms. Spatial attention helps focus on the most informative spatial regions within feature maps. This is crucial for accurately localizing fine details relevant for the upscaling process. Channel attention recalibrates the weights of different feature channels. This allows SRRAM to emphasize channels carrying highly relevant textural and structural information for reconstruction. By simultaneously employing spatial and channel attention, it guides the network to learn both where to focus (spatially) and which types of features to prioritize (channels), enabling the capture of complex image structures necessary for high-fidelity SR.

SRRAM uses Adam optimizer known for its adaptive learning rate adjustment and L1 loss function. It produces sharper looking SR output. Designing computationally efficient attention modules suitable for real-world deployment on less powerful devices is an important direction for SRRAM-like models.

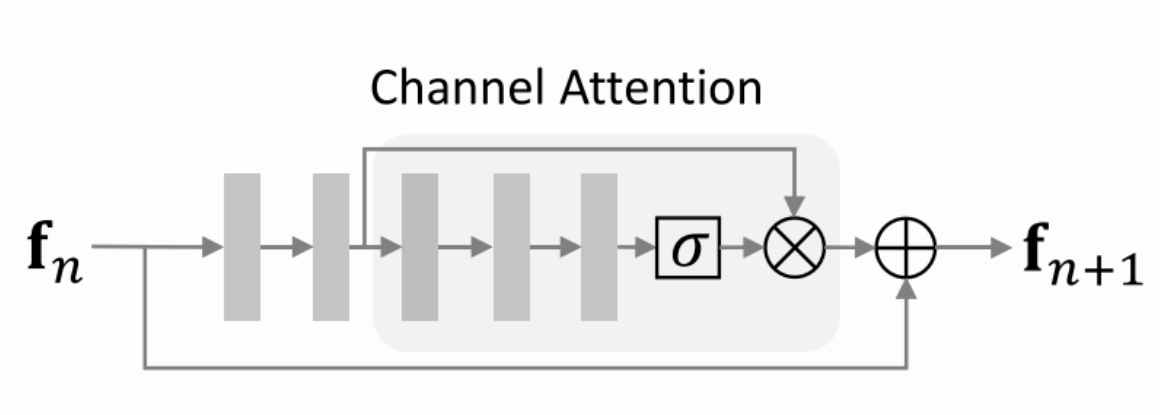


Figure2.8
Simple representation of network architecture for RCAN and SRRAM (Anwar et al., 2019).

The SR networks discussed so far consider Bicubic degradations. In real world scenario, it is feasible to have multiple degradations occurring simultaneously. To deal with this, multiple degradation handling network methods are proposed. Two of such methods are Zero-Shot Super Resolution (ZSSR) and Super Resolution Network for Multiple Degradations (SRMD). The network learn relationship between low resolution test image and ground truth test image. Then same network is used to predict SR image while using test image as an input.

Thus, the network doesn't need training images for a given degradation and network can learn an image specific network of feature map on the fly during inference. ZSSR uses a very simple network architecture trained using down-sampled version of test image and learns residue image using L1 norm. This allows ZSSR to adapt to the specific image and degradation at hand during the inference phase.

SRMD takes a concatenated low-resolution image and related degradation maps. It tackles the problem by requiring both the degraded LR image and a 'degradation map' as input. This map provides information about the spatial distribution and type of degradation the image has suffered. The degradation map allows the network to learn a more informed representation tailored to the specific distortions present, improving the reconstruction process. It uses deeper network with convolutional layers followed by sequences of blocks, enabling it to model complex mappings between degraded and clean images. The network architecture of SRMD consists of multiple layers such as cascade of convolutional layers followed by sequence of Conv (Simonyan and Zisserman, 2014), ReLU and Batch normalization layer. SRMD learns high-resolution images instead of residue of images. SRMD is known for its unique capability to handle multiple degradations.

ZSSR offers exceptional flexibility on unseen image degradations. SRMD might achieve superior results if degradation types are known beforehand and suitable degradation maps can be generated. ZSSR's test-time adaptation may come with increased computational overhead compared to SRMD's pre-trained approach.

This brings us to end of discussion for various deep learning based Super Resolution methods.

2.3 An overview of deep convolutional neural network work for Super Resolution

I briefly discussed about various deep learning-based methods for SR. In this section, I would explain some of the recent and profound work completed for SR using deep convolutional neural networks. These works have set future direction for SR work in computer vision community. Authors of (Dong et al., 2016) considers a CNN which directly learns a complete mapping between low-resolution and high-resolution images. The proposal is to create SR pipeline just from learning with very little pre or post processing beyond the optimization. This proposed model is called Super

Resolution Convolutional Neural Network (SRCNN). This model shows that restoration quality of network can be improved with larger and diverse set of datasets and also with a deeper larger network. This model shows that deep learning is really useful in SR. (Dong et al., 2016) continue to work on SRCNN. The performance of SRCNN is further improved by using larger filter size in non-linear mapping layers, addition of three colour channels and inclusion of more test images. In the improved SRCNN model, the task was to learn a mapping which consist of three operations:

- Patch extraction and representation: This involves extraction of overlapping patches from low-resolution image and then represent every patch as a high-dimensional vector. The vectors consist of feature maps. Extracted patches are represented as a set of pre-trained bases. This enhancement introduces the concept of representing each patch as a combination of pre-trained bases (similar to a simplified dictionary learned from image data). This potentially injects a degree of prior knowledge into the representation process. The next step is including optimization of these bases into the optimization of network. By including the optimization of these bases directly into the network's learning process, the representational bases themselves are adapted alongside the core CNN weights, potentially leading to a more effective representation for the super-resolution task.
- Non-linear mapping: This involves mapping every high-dimensional vector onto another high-dimensional vector. These vectors consist of set of feature maps. The goal is to learn an optimal non-linear mapping specifically for reconstructing high-resolution patches from their compressed features. Finally, ReLU is applied on filter response. ReLU aids in learning more complex mappings and promotes efficient gradient flow.
- Reconstruction: This involves aggregation of first two steps high-resolution feature-wise representation to generate final high-resolution image which is expected to be similar to ground truth. The improved SRCNN incorporates the idea that the predicted HR value of a

pixel should consider not just its own patch but also information from neighbouring patches within a spatial region.

All these three methods lead to form a convolution neural network where the task is to optimize filtering weights and biases. The figure represents above mentioned three operations. First convolutional layer extracts set of feature maps, second layer maps extracted feature maps nonlinearly to high-resolution patch representation and last layer adds predictions within a spatial circle to produce final high-resolution image.

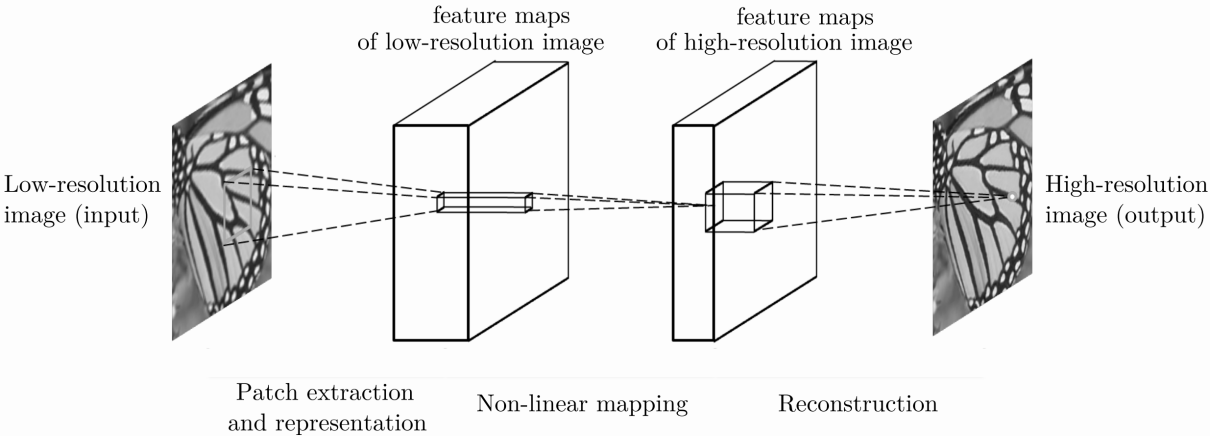


Figure2.9
Three operations of improved SRCNN methods for Super Resolution (Dong et al., 2016)

This improved SRCNN also uses Mean Squared Error (MSE) loss function. The loss is minimized using stochastic gradient descent with stand backpropagation, this standard optimization algorithm is used to adjust model parameters (both the regular CNN weights and potentially the pre-trained bases) iteratively based on calculated gradients. All convolution layers have no padding to avoid border effects during training and simplify the mapping that needs to be learned.

The experiments conducted by (Dong et al., 2016) have shown keeping a lightweight structure can result in superior performance of improved SRCNN model. Even a relatively shallow network can achieve strong results through enhancements to its feature representation and optimization strategy. This highlights that the most significant SR gains don't necessarily come from simply stacking more layers. It also highlights that 'the deeper the better' is not always true for all deep convolutional neural network models for SR. The experiments show that going deeper i.e. more layers in improved SRCNN makes it hard to set appropriate learning rates to guarantee convergence. Even in case of convergence, there is a chance that network may fall into a bad local minimum even with enough training time. There are practical and training complexities that arise for SR tasks, such as convergence risks because of deeper variants of the improved SRCNN proved more difficult to train, it becomes harder to find suitable learning rates that ensure convergence, and even then, the risk of the model getting stuck in a suboptimal solution remains a concern, and diminishing returns explained in experiments which indicate that beyond a certain depth, accuracy either plateaus ("saturation") or even starts to decline ("desaturation"). This suggests that simply adding more layers may not improve the model's capacity to extract and utilize relevant features for super-resolution.

Improved SRCNN model doesn't have pooling or fully-connected layer which results in it being sensitive to initialization parameters and learning rate. The experiments also show that improper increase of depths leads to accuracy saturations or accuracy desaturation. This highlights the need for meticulous hyperparameter tuning and potentially advanced optimization strategies to ensure stable training. This work paves the way for various experiments in SR space by exploring more filters, different training strategies and different upscaling factors.

Authors of (Kim et al., 2016) found that work completed by (Dong et al., 2016) has limitations for SR problem space. They noticed that work completed by (Dong et al., 2016) relies

on the context of small image regions which restricts their ability to leverage long-range dependencies and broader contextual information that could aid in reconstructing finer textures and details. Not only this observation but they also noticed training converges too slowly with deeper layers and network just work for single scale, this limits their flexibility for real-world scenarios with varying scaling needs.

(Kim et al., 2016) proposed a new highly accurate SR method based on very deep convolutional network. They solve the problem of deep network converging slowly by boosting converging rate with high learning rate and spreading contextual information over very large image regions. This introduced exploding gradients. They solve exploding gradients with residual learning and gradient clipping while also make network work with multi scale SR. They called this model Very Deep Super Resolution (VDSR).

A common problem with very deep networks is to predict dense output of continuously reducing feature map size every time a convolution operation is performed. In context of SR, for pixel near image outer areas (boundary) are never exploited to full extent and thus many SR methods crop the result image. This doesn't work for cases where output image requirement has surround region to be very big.

To resolve this issues, authors of (Kim et al., 2016) pad zeros before convolution to keep size of all feature maps same. This turns out surprisingly well and output image pixels near image outer areas are also correctly predicted. This simple but effective technique ensures that even boundary pixels have sufficient contextual information for accurate reconstruction. After image details are correctly predicted, they are added back to the input low-resolution image to create final high-resolution image. An elegant process that leverages both the residual details and the information contained within the upscaled input.

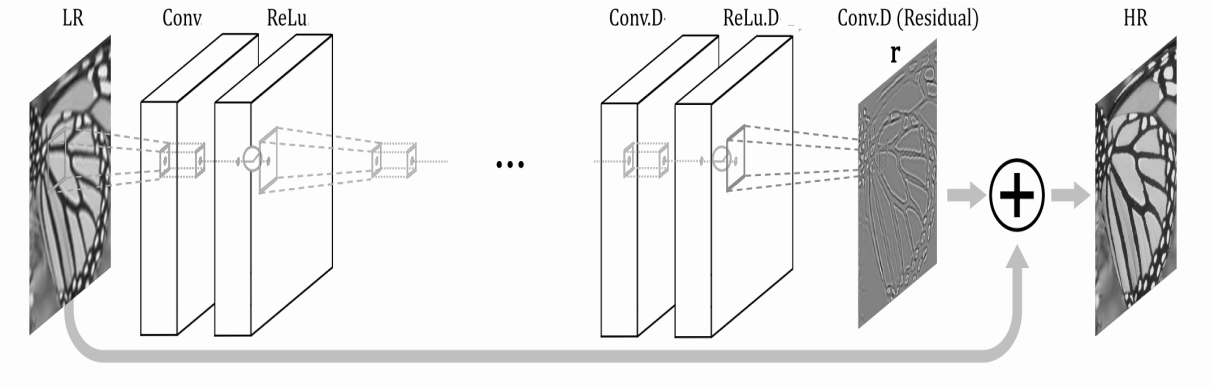


Figure2.10
Network structure of VDSR (Kim et al., 2016).

The experiments conducted by (Kim et al., 2016) for VDSR has shown that by incorporating residual learning, extremely high learning rate with gradient clipping, it is feasible to converge very deep CNN for SR. Finally, with residual learning, gradient management, and careful design choices, deep CNNs can be successfully trained for super-resolution tasks. VDSR features a 20-layer network, a notable jump from the relatively shallow models of the time. It can also be used in other image restoration tasks such as compression artefact removal and denoising, highlighting the broad utility of its approach.

The work proposed by (Dong et al., 2016; Kim et al., 2016) leads suggested improvement in existing deep learning based CNN architectures (Simonyan and Zisserman, 2014) or leads to generation of new deep learning based CNN architecture for SR (Anwar et al., 2019) and face detection and face alignments (Zhang et al., 2016). The work also lead to creation of new datasets (Cao et al., 2018) because of increasing demand from computer vision community to analyse larger more complex datasets for feature extraction.

This brings us to end of discussion for deep convolutional neural network work for Super Resolution methods.

2.4 A review of recent GAN based Super Resolution methods

The process of SR is still considered a very challenging problem in computer vision and image processing community. The SR process becomes even more challenging to when applied on face images or face feature extraction. There are well known general problems in SR space such as:

- Ill-posed inverse problem – There exist multiple solutions for a given low-resolution image, thus it requires reliable prior information to constrain solution space. For each low-resolution image (Johnson et al., 2016), there exist multiple high-resolution images that could have generated it, the process discards fine details and recovering these lost details is inherently ambiguous. This becomes more extreme with high upscaling factors; the network has to essentially 'invent' missing information without strong guidance. For large factors (upwards of 4x), minute details of the high-resolution image may have little or no existence in its low-resolution version.
- Higher upscaling factors leads to increase in complexity – With higher up-scaling factors, recovery of missing details become complex and leads to reproduction of wrong data. Humans are highly attuned to facial details. Even minor inaccuracies significantly affect the perceived identity and naturalness of the reconstructed face. It's not just about pixel-level accuracy, the model needs to hallucinate details (e.g., wrinkles, subtle textures) in a way that looks believable and natural for a human face. Also, recovery of finer texture details become very challenging.
- Complexity in assessment of output quality – It is not straightforward to assess quality of output. Even with high Peak Signal to Noise Ratio (PSNR), up-sampled images often lack high frequency details and are visually unsatisfying. An overly smooth HR image might have a high PSNR but appear unrealistic. This sums up known fact in SR literature that

well-known quantitative measures do not encapsulate perceptual soundness of generated high-resolution output. Ideally, we want super-resolved images that are both pixel-wise accurate and visually convincing, with realistic textures and details.

Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) based SR models aims to solve known SR space problems without complexity and provide much greater upscaling factors while also preserving finer texture details, not only at central part of image but also at outer boundary areas. GAN encourage the network (Wang et al., 2018a) to favour solution which look similar to natural images.

The core of GANs is the competition between a generator (creating SR images) and a discriminator (trying to distinguish real HR images from fakes). This trains the generator to produce outputs that are indistinguishable from the distribution of real HR images, introducing a powerful prior and reducing the ambiguity of the SR problem.

GANs can be trained with perceptual loss functions that focus on making the super-resolved image look realistic and visually pleasing, even if it means some deviation from pixel-wise exactness compared to the HR ground truth. This aligns better with human assessment of quality. Lastly due to their focus on learning the distribution of natural images, GANs become remarkably adept at 'filling in' missing details in a convincing manner. This is particularly beneficial for recovering fine facial features that might not be directly deducible from the LR input alone.

In this section, I will discuss various State-Of-The-Art (SOTA) GAN model for SR image process.

2.4.1 Super Resolution Generative Adversarial Network (SRGAN)

SRGAN represents a major breakthrough in super-resolution for its focus on photorealism and the introduction of perceptual quality metrics. This is the first framework known for inferring photo-realistic natural images at up-sampling factors of 4x (Ledig et al., 2017). (Ledig et al., 2017)

work highlighted that while minimizing MSE which in turn maximizes PSNR is a common measure to evaluate SR methods, the ability of these metrics to capture relevant differences at high texture details is very limited.

SRGAN's key contribution was the integration of perceptual losses operating on feature representations extracted from pre-trained networks. SRGAN uses perceptual loss functions (content loss, adversarial loss) using high level feature maps of Visual Geometry Group (VGG) network with a discriminator to create perceptually hard to distinguish SR images. Their experiments give more value to Mean Opinion Score (MOS) testing compared to PSNR or SSIM metrics. SRGAN consists of a multi-task loss formulation with these parts:

- MSE loss to encode pixel-wise similarity.
- A perceptual similarity metric defined over high-level image representation (feature maps) in terms of distance metric. This is the new perceptual content loss. Based on high-level feature maps from the VGG network, this loss encourages the SR image to be perceptually similar to the ground truth HR image, not just pixel-by-pixel identical.
- An adversarial loss to balance min-max game between discriminator and generator (this is a standard GAN objective). This pushes the network towards learning the complex distribution of textures and details that characterize realistic images.

The generator network has residual blocks with identical layout. The generator leverages residual blocks with skip connections to aid training of deeper architectures. It contains two convolutional layers (3x3 kernels) and feature maps (64). This is followed by batch normalization and ParametricReLU activation function, aims to improve model fitting. The discriminator uses LeakyReLU activation and avoids max-pooling to help preserve more spatial information for discrimination between real/fake HR images. throughout the network.

It contains eight convolutional layers with an increasing number of 3x3 filter kernel (increased by factor of 2 from 64 to 512). Each time number of features are doubled, strided convolutions are used to reduce image resolution, increasing its receptive field with fewer parameters compared to max-pooling. The final 512 feature maps are followed by two dense layers and a final sigmoid activation function to obtain a probability. This is shown in figure X. 'n' represent number of feature maps and 's' represent stride.

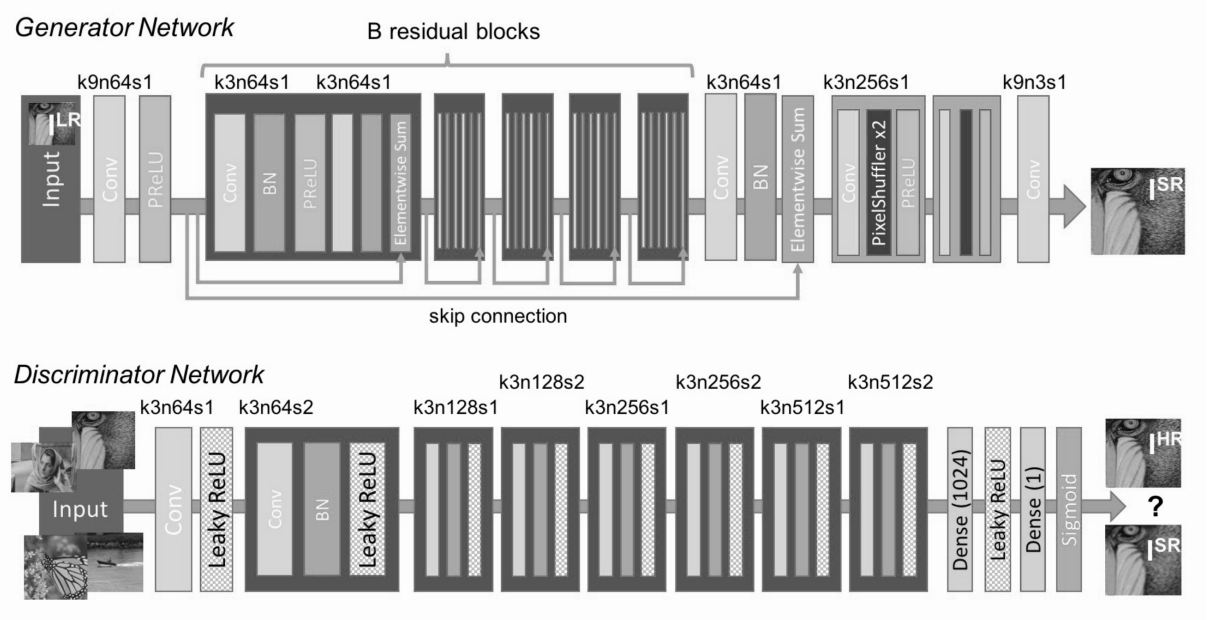


Figure2.11
SRGAN structure showing generator and discriminator (Ledig et al., 2017).

(Ledig et al., 2017) sets first SOTA for image SR with 4x upscaling when measured on standard PSNR and SSIM with 16 blocks deep ResNet optimized for MSE. Further, they propose SRGAN by replacing MSE-based content loss with a loss calculated on feature maps of VGG and they proved with extensive MOS testing that SRGAN produces most photo-realistic SR images.

The modularity of SRGAN, with its separation of content and adversarial losses, provided a rich space for experimentation with different perceptual loss functions, network architectures, and discriminator designs. This flexibility makes it a valuable starting point for this research work.

Because of various tweaking options provided by SRGAN, this makes it ideal candidate to tweak, extend and create new GAN models for SR image space. SRGAN achieved state-of-the-art performance at the time, not only in PSNR/SSIM but, critically, by significantly improving perceived image quality through its emphasis on perceptual realism. This thesis will use SRGAN as base with various feature extractors to test SOTA in SR image space.

2.4.2 Super Resolution Perceptual Generative Adversarial Network (SRPGAN)

While SRGAN's use of VGG features improved photorealism, the argument is that VGG, primarily designed for classification, may not be the most discriminative feature extractor for capturing the nuances of super-resolved image quality. Authors of (Wu et al., 2017) proposed a GAN based model to recover high frequency details from high-resolution images. The proposal involves a robust perceptual loss based on the discriminator combined with Charbonnier loss function to create content loss and merge it with adversarial loss plus proposed perceptual loss.

SRPGAN innovates by defining its perceptual loss based on intermediate feature maps extracted directly from its own discriminator network. It reuses existing computation during adversarial training, avoiding the overhead of a separate VGG network. This gives SRPGAN ability to construct SR images with great details and sharp edges.

The model replaces batch normalization layer with instance normalization layer. The shift is that instance normalization normalizes features within a single image instance, rather than across a batch. This potentially be more suitable for handling variations within individual images during the super-resolution process.

(Wu et al., 2017) notes that perceptual loss in SRGAN (Ledig et al., 2017) is based on VGG classification network and can’t capture required high frequency details in SR images tasks, it also introduces extra computation overhead. This leads to design of a more robust perceptual loss

function. In SRPGAN, traditional pooling layer is replaced with stride-convolutional layer, this helps in preserving finer-grained spatial information for the discriminator's analysis.

Instance normalization replaces traditional batch normalization in every layer of generator. Instance normalization apply normalization on s single image instead of set of images. For perceptual loss, it uses intermediate features of discriminator to build loss function, a multi-part loss function. It still includes an adversarial term and a content loss but replaces MSE with the Charbonnier loss (a more robust variant of L1) and the new discriminator-based perceptual loss. This reduces computational overhead by reusing extracted features from discriminator compared to a loss function build on VGG. SRPGAN uses ADAM optimizer for both generator and discriminator.

From the experiments completed by (Wu et al., 2017), it is clearly visible that SRPGAN method reconstructs better fine details compared to SRGAN. It is also less intensive computationally compared to SRGAN because it doesn’t need an extra VGG classification net to build perceptual loss. Its focus on efficiency and its refined perceptual loss strategy highlight the continued research into optimizing the trade-off between computational cost and super-resolution realism.

2.4.3 Enhanced Super Resolution Generative Adversarial Network (ESRGAN)

Authors of (Wang et al., 2018a) studied SRGAN (Ledig et al., 2017) and they found that it leaves unpleasant artefacts. The study focused on three key areas of SRGAN. First is network architecture, second is adversarial loss and third is perceptual loss. They improved each of these three aspects of SRGAN to develop ESRGAN and improved overall perceptual quality. ESRGAN builds upon the success of SRGAN and SRPGAN, recognizing areas for improvement. It targets a more refined and visually satisfying SR output through modifications in three primary aspects.

In a similar approach to SRPGAN, they removed batch normalization and introduced Residual-in-Residual Dense Block (RRDB) with smaller initializations and residual scaling to facilitate training a very deep network. These deeper nested structures with densely connected layers and residual scaling promote better feature representation for complex mappings.

Also, they introduced relativistic average GAN (RaGAN) to discriminator. RaGAN allows discriminator to evaluate and decide whether one image is more realistic than another image. This allows discriminator to predict relative value instead of absolute value. Instead of directly predicting whether an image is 'real' or 'fake', RaGANs focus on relative realism. The discriminator learns to rank one image as relatively more realistic than another.

ESRGAN provides better brightness consistency, sharper edges, texture recovery and more visually pleasing results by using the feature map before activation in perceptual loss. It extracts features for loss calculation before the activation layer in the VGG network. This change aims to capture more unprocessed, distribution-like characteristics of the images. Brightness consistency, combined with the other improvements, leads to outputs with better overall brightness balance and less variation in brightness compared to predecessors.

It also used image interpolation to directly interpolate images pixel by pixel, providing more variations in the input image domain and contributing to a more well-rounded learning process. The network architecture is shown in figure X. Here, β is residual scaling parameter.

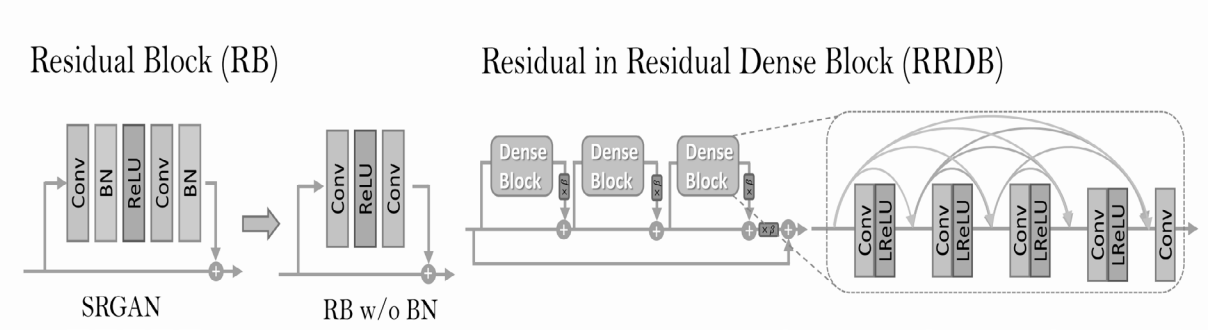


Figure2.12
ESRGAN (Wang et al., 2018a) structure showing changes compared to SRGAN.

From the experiments completed by (Wang et al., 2018a), it is clearly visible that ESRGAN produces sharper, more natural looking with better texture SR images compared to SRGAN. It also produces better more detailed structures in output SR images. In both qualitative (visual examples) and quantitative (perceptual metrics) evaluations, ESRGAN outperforms earlier SRGANs. Images generated by ESRGAN boast sharper edges, richer and more believable textures, and an overall more realistic feel. It shows notable improvement in the recovery of complex structural details in super-resolved images, emphasizing the value of its network and discriminator enhancements.

2.4.4 Progressive Generative Adversarial Network for Super Resolution (ProGANSR)

Authors of (Wang et al., 2018b) has proposed a progressive method for GAN architecture and training. Their proposal is to up-sample an image in intermediate steps (2x upscaling) where learning process is organised from low easy level to high hard level. The approach uses this progressive design to multi-scale to very high 8x upscaling while increasing reconstruction quality of generated SR image. It introduces a fundamental shift in GAN-based SR by tackling the problem progressively and focusing on computational efficiency.

To make it run faster, the model uses dense compression units and simplified network layout of faster information propagation through network. To make it efficient at very high upscaling, the

model uses an asymmetrical matching pyramid structure with more layers at lower level. The complete end to end progressive design allows the model to have a multi-scale generator with a unified discriminator in a single training. Instead of directly training on the final upscaling factor (e.g., 8x), it incrementally increases the resolution during training.

It starts with a low-resolution image and gradually adds generator and discriminator layers that double the resolution in each step (2x, 4x, 8x). This breaks down the complex SR problem into smaller, easier tasks. Early stages focus on learning coarser features, which then provide a solid foundation for finer detail reconstruction in later stages.

Instead of using traditional loss function (content loss plus perceptual loss), ProGANSR operates on residual output of each scale and train progressively. It uses curriculum-based learning to further improve run time performance and reduction in training time.

The proposed model is very lean on computing resources and run very well, thus making it suitable for real life applications. It leverages dense compression units within its architecture for efficient feature flow, reducing computational and memory overhead. To handle the greater complexity of large upscaling factors, it employs more layers in the lower-resolution stages.

This mirrors how humans might tackle the problem, focusing more attention on coarser structures before refining details. It doesn't train on the absolute output of each scale but focuses on the residual difference compared to the lower-resolution prediction. The staged training aligns with 'curriculum learning' principles, where easier tasks are presented before increasing difficulty. This boost overall training stability and efficiency. Compared to upscaling factor of 4x in SRGAN and EnhanceNet (Anwar et al., 2019), ProGANSR upscales to 8x factor.

A more advance research has been presented by (Zhong et al., 2019) for face image SR mixing up SRPGAN, ESRGAN and ProGANSR. It uses Residual Dense Block (RDB) with RaGAN, Inception architecture and feature maps before activation with progressive architecture

and training. This network shows excellent performance in reconstruction of low-resolution face images under large scaling factors of 4x and 8x.

2.4.5 Cycle in Cycle Generative Adversarial Network (CinCGAN)

Authors of (Yuan et al., 2018) took a unique case of image SR problem where low-resolution and high-resolution pairs, along with down-sampling process is not available. In this case, low resolution input image is further degraded by noise and blurring. This type of setup makes kernel estimation and supervised learning impossible. They proposed a Cycle in Cycle GAN with three steps to tackle the problem.

First step is to map noisy and blurry image to a noise-free low-resolution space. It aims to map the degraded LR image into a cleaner LR space. This presumably helps simplify the subsequent upscaling problem. The second step is to up-sample the image with a pre-trained deep model, the noise-reduced and deblurred image is upscaled using a pre-trained deep super-resolution model. The final and third step is to fine-tune first two steps in a complete manner to achieve high-resolution image. Critically, the first two stages are fine-tuned jointly. This end-to-end optimization adapts both the initial denoising/deblurring step and the SR network parameters to better work in conjunction, even though they weren't explicitly trained on corresponding pairs.

The network architecture involves unpaired image to image translation. It leverages CycleGAN-like architectures for its core mappings. Crucially, this allows it to learn without requiring paired LR-HR examples. It uses convolution layers for their feature extraction and transformation properties with LeakyReLU layer and batch normalization layer. The "Cycle in Cycle" aspect likely refers to imposing losses that encourage not just LR to HR generation but some form of reconstruction back to the original degraded LR domain, adding a regularization effect to ensure the mappings don't drift too far from the input distribution.

Experiments performed by (Yuan et al., 2018) shows that FSRCNN (Anwar et al., 2019) and EDSR (Anwar et al., 2019) doesn't work well if training process is unaware about blur and noises compared to CinCGAN, this emphasizes the need for such techniques that can adapt and refine even pre-existing SR methods to match real data.. The problem setting addressed by CinCGAN is closer to many real-world image restoration needs, where high-quality reference images are simply unavailable.

2.4.6 Spatial Attention Generative Adversarial Network (SAGAN)

Authors of (Zhang et al., 2018) worked on creating GAN model for face attributes by introducing spatial attention mechanism into GAN. They showcased that existing GAN based model while work for face attributes editing, they inevitably change irrelevant regions as well, unintended modifications to areas outside of the specific attribute being targeted. In SAGAN, spatial attention mechanism allows one to select specific prior parts and ignore the rest of image for faster and accurate processing. SAGAN's core innovation is the integration of a spatial attention mechanism to address this issue.

Spatial attention allows the model to dynamically focus on specific regions of the face image related to the desired attribute change, promoting more accurate editing. Attention guides the model to avoid making changes to irrelevant portions of the image, reducing unwanted artifacts and potentially improving the overall naturalness of the result.

Generator is given an attribute manipulation network for face attribute editing and a spatial attention network to localize editing to an attribute specific region. The discriminator does the usual work of identifying generated images from real ones. The additional task for discriminator is to classify the face attributes, further encouraging the generator to create outputs that align with the desired attribute changes.

SAGAN is designed to use a conditional input signal representing the desired attribute modification to guide the generator. It uses a single generator with conditional signal as attribute. It uses Adam optimizer. For every run, the generator is updated once white discriminator is updated three times. The update ratio of generator vs. discriminator (1:3 in the description) is an interesting experimental decision. Training GANs often involves careful balancing to find stable and effective optimization regimes.

Experiments performed by (Zhang et al., 2018) shows that SAGAN performs better than existing face attribute editing GAN models such as CycleGAN (Singh and Raza, 2020), StarGAN and ResGAN because of spatial attention mechanism. It's emphasis on spatial attention lead to more controlled edits that precisely target the desired facial attribute while preserving other aspects of the face. Spatial attention mechanisms enable greater control and precision in image synthesis tasks.

2.4.7 Real Time Super Resolution Generative Adversarial Network (RTSRGAN)

Authors of (Hu et al., 2019) worked on creating a GAN model for SR at client devices. They noticed that existing GAN based models (such as ESRGAN (Wang et al., 2018a)) are impractical because of high requirement of computing power. Such computing power are generally not available at client-side devices which can run a SR real time video playback. They noticed that while ESPCN (Anwar et al., 2019) has best real time performance, the offered perceptual quality is not good. The RTSRGAN addresses these critical practical issues in super-resolution.

To solve real-time demand of SR video playback, (Hu et al., 2019) decided to take best features from ESRGAN and ESPCN. They decided to exploit the advantages of ESRGAN for perceptual image quality and ESPCN efficiency for real-time performance.

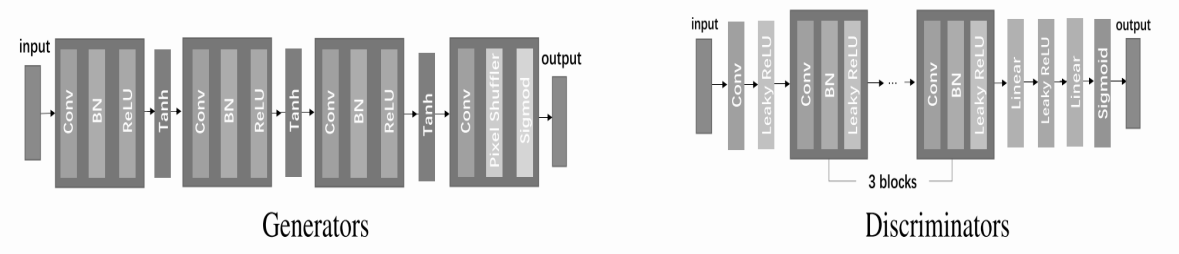


Figure2.13
RTSRGAN (Hu et al., 2019) structure showing generator and discriminator.

In RTSRGAN generator, batch normalization and ReLU are added to ESPCN. Number of layers are reduced in discriminator for ESRGAN. Batch normalization and ReLU layers give generator to have a very high generation ability while also accelerating reconstruction, are strategically added to enhance the generator's representational capacity and accelerate the reconstruction process. The ESRGAN discriminator's depth is reduced, decreasing its computational cost while retaining its core ability to distinguish fine details. The generator and discriminator are put to training at same time but just generator is applied during test and real-time use for efficiency.

Experiments shows that RTSRGAN has very good efficiency and low resource requirement while maintaining a very high reconstruction speed for SR images in real-time video playback. Comparisons on metrics like FLOPs (Floating-Point Operations Per Second), inference time per frame, and overall processing speed demonstrates its edge against ESRGAN while potentially surpassing ESPCN. Even with its efficiency focus, it's important that RTSRGAN holds its own in terms of PSNR/SSIM on standard SR datasets.

2.4.8 Face Completion Super Resolution Generative Adversarial Network (FCSR-GAN)

Authors of (Cai et al., 2020) noticed that in case of video surveillance, face images have variation because of low-resolution, occlusion and partial occlusions (objects blocking portions of the face). This severely impacts both the clarity and potential for accurate identification. They also noticed that existing face image recovery methods can handle just one type of variation per model, often focus on a single type of degradation at a time, either super-resolution or occlusion completion and inpainting. This isn't ideal for real-world cases where both problems co-exist.

Thus to overcome these limitations, (Cai et al., 2020) proposed a GAN model using multi-task learning to perform joint face image SR and face image completion at same time. The task is to recover or construct a high-resolution occlusion free face image from low-resolution face image with occlusion.

To achieve this, generator performs face SR and face completion simultaneously, targeting to recover a high-resolution face image without occlusion from input image of low-resolution with occlusion. The generator's core task is to produce a high-resolution, occlusion-free facial image starting from its degraded low-resolution, occluded input. The generator might contain modules or branches specialized for super-resolution and inpainting. How these modules interact and potentially share information is a crucial design aspect often experimented in various research.

The discriminator employs six losses. First is an adversarial loss for differentiating between real face images and generated face images remains central to the GAN framework. However, FCSR-GAN takes it further. Second is a pixel loss for reconstruction of high-resolution face image without occlusion, encourage pixel-level accuracy of the reconstructed HR occlusion-free image. This is essential but risks overly smooth outputs without additional constraints. Third is a

perceptual loss aiming to get photorealistic texture or fine details, encourages the generation of realistic textures and fine details, minimizing the "artificial" appearance of generated faces.

Fourth is a smooth loss for penalizing colour space distortion at the outer boundary area, an intriguing addition! this loss aims to ensure a seamless transition between the recovered facial areas and the original non-occluded regions, preventing abrupt colour or texture changes that would reveal the reconstruction process. Fifth is a style loss for keeping the style between occlusion free facial area and recovered facial area, this addresses a subtle but important aspect of realism. Finally, sixth loss is a face prior loss to get a facial topology structure, enforces knowledge about typical facial structures and proportions, providing a strong regularizing effect and ensuring that the generated output adheres to plausible face anatomy.

Experiments performed by (Cai et al., 2020) shows that FCSR-GAN performs much better in leveraging a multi-task learning and delivering solid qualitative output in SR face image. It's multi-task nature, with all these losses, requires careful balancing during training to ensure that all aspects of the image reconstruction progress effectively. It also shows that even when ground truth occlusion free high-resolution face images are not available, the recovered high-resolution occlusion free face image looks perceptually very pleasing.

FCSR-GAN outperform models that separately handle SR or completion, highlighting the benefits of its joint approach. The ability to generate high-resolution, occlusion-free faces that appear both natural and consistent with the non-occluded parts of the input is great. The results also show that FCSR-GAN can perform easily perform 8x upscaling with great looking visually pleasing output SR images.

2.4.9 Enhanced Super Resolution Generative Adversarial Network+ (ESRGAN+)

Authors of (Rakotonirina and Rasoanaivo, 2020) noticed that there is a gap between ground truth images and SR images generated by ESRGAN and they proposed an enhanced version of

ESRGAN to close the gap. The core goal is minimizing the remaining perceptual discrepancies between super-resolved images and real HR ground truth, leading to more realistic and finer-grained results. They noticed that perceptual quality of ESRGAN (Wang et al., 2018a) can be further improved by introducing a new block Residual-in-Residual Dense Residual Block (RRDRB) which has higher capacity compared to RRDB block of ESRGAN.

Further, they added noise input to generator in the network to take advantage of stochastic variation by adding Gaussian noise to output of each residual dense block. ESRGAN network has RRDB block. The rationale behind injects noise (often Gaussian) into the output of each residual dense block within the generator is two-fold; one the added noise introduces random perturbations during training. This forces the network to continually adapt, preventing it from getting stuck in suboptimal solutions. Second, in image generation, sometimes overly deterministic paths lead to overly smooth and somewhat "artificial" results. Noise breaks this determinism during training, potentially leading to more organic-looking textural nuances in the output.

It is further enhanced in ESRGAN+ by addition of an additional level of residual learning inside the dense block to support network capacity without increasing complexity. The extra layers give the network more capacity to learn and transform features, potentially leading to a more nuanced super-resolution process. Residual connections within the dense block encourage reuse of previously extracted features, while additional layers also create pathways for finding entirely new features. By continuing this trend, a residual is added every two layers in each dense block. With this new addition, it enables both re-use of feature and capacity to find new features.

While ESRGAN significantly raised the bar in SRGAN-based realism, ESRGAN+ highlights that there's still subtle room for improvement in matching the complex distributions of real image textures. It embodies the continued trend in SRGAN development of gradually refining architectural components for continuous enhancements in output quality. Experiments conducted

by (Rakotonirina and Rasoanaivo, 2020) shows that reconstructed SR images have more detailed structures and are less distinguishable from input ground-truth low-resolution images. Texture details are preserved well.

2.4.10 Newly developed Generative Adversarial Network models for Super Resolution

So far, we have discussed some SOTA GAN models for SR image processing and generation. I would also like to mention recent GAN models for SR image. These models are newly published and may not be in use extensively but they have outperformed SOTA methods in SR space. I will briefly touch over such GAN models.

Authors of (Zhang et al., 2020) noticed that lack of prior knowledge and small receptive field of existing CNN are not yet addressed for SR problem space. Many SR approaches rely primarily on learning mappings directly from the data, potentially missing out on valuable prior knowledge about image structures, textures, or domain-specific characteristics that could aid the reconstruction process. The local nature of convolutional operations means that standard CNNs often struggle to capture long-range dependencies within an image.

This is crucial in SR, where the reconstruction of a high-resolution detail might require contextual information from distant regions of the low-resolution input. They propose Segmented Prior Self Attention Generative Adversarial Network (SPSAGAN) to mix priors and feature space into a unified framework. The term "segmented" explains different types of priors being used depending on local image content or the specific image domain. This enables the network to consider relationships between distant pixels for better detail reconstruction.

The network uses Residual-in-Residual Sparse Block (RRSB) to get even more performance with reduction in computation cost. Sparsity can improve computational efficiency and reduce overfitting risks. It aims to explicitly incorporate prior knowledge and guide the super-

resolution process alongside the standard CNN feature extraction pipeline. Experiments shows SPSAGAN generates more realistic and visually pleasing textures in SR image space.

Authors of (Ma et al., 2020) notices that there are always undesired structural distortions in generated SR image. Even when a model produces realistic-looking outputs in terms of textures and fine details, it might still introduce subtle structural distortions. These might manifest as slightly warped lines, deformed shapes, or distortions in the overall spatial relationships between image elements. They propose a structure persevering SR method to address this structural distortion issue and named it Structure Preserving Super Resolution Generative Adversarial Network (SPSRGAN). It addresses the subtle yet crucial issue of structural distortion in super-resolution outputs, using gradient-based approach, and keep emphasis on maintaining perceptual quality alongside structural accuracy.

The proposed method raises above issue while generating perceptually pleasant details in generated SR image. The method uses high-resolution gradient map to provide structural priors and uses a gradient loss to impose a second order restriction on generated SR image. Gradient maps inherently capture edges, boundaries, and the directionality of changes in intensity, the core elements that define image structure. Gradients can be computed at different scales. This provides a form of structural information independent of the initial degradation faced by the LR input. Gradients measure the change in image intensity.

Thus, the gradient loss effectively imposes a constraint on the local patterns of change allowed in the output, preserving structural integrity. This helps make method model-agnostic, an important advantage of the gradient loss. It likely be applied in conjunction with various SRGAN architectures. It doesn't require fundamental modifications to the SR generator itself, but acts as a supplemental constraint during training. Experiments shows SPSRGAN generates superior natural looking SR images while keeping structures intact.

Authors of (Zhong and Zhou, 2020) noticed unpleasant artefacts and distortions in generated SR images. This finding acknowledges a common issue in GAN based SR, the tendency to produce artifacts and distortions that detract from the overall quality of the generated image, even if it has areas of plausible texture and detail. These issues may stem from instabilities in training and the complex mappings learned by the generator.

They proposed to tackle these problems using Latent Space Regularization Generative Adversarial Network (LSRGAN) model to optimize supervised GAN. It explicitly introduces the concept of optimizing the mappings learned by the GAN, not just in the traditional image space but also in a latent space. This latent space can be seen as a domain where the generator has more freedom to explore variations and combinations of features. The aim is to impose constraints or smoothness on those mappings within the latent space.

Regularization can guide the generator towards learning a well-behaved mapping that is less likely to produce abrupt, unnatural transitions as it produces the final output, reducing the likelihood of artifacts. After this, they explicitly apply Lipschitz Continuity Condition (LCC) to regularize supervised GAN in order to map image space to a new optimal latent space while augmenting underlying GAN as a coupling component. The LCC is a mathematical criterion often used to ensure stability and controllability of functions.

The intuition is that this constraint will discourage the generator from learning mappings that have sudden jumps or overly steep changes, which could manifest as distortions or artifacts in the output image. Finally, generator is asked to converge to an ideal mapping that can generate image sample which are faithful to target images. To sum it up, LSRGAN treats a standard GAN architecture (generator and discriminator) as one piece of the puzzle.

The core focus is on transforming and regularizing its learned representations. It seeks to find an 'optimal' latent space where the generator's mapping to the final image space is well-

behaved and fulfils the LCC. Finally, the goal is for the generator to ultimately create outputs that are highly faithful to the ground truth high-resolution targets even after the latent space transformations and optimization. Experimental Evidence Experiments shows this model outperform existing SOTA methods in SR image space.

Authors of (Ditria et al., 2020) noticed that existing conditional GAN models are limited because of pre-defined and fixed class level semantic attributes, a common limitation in conditional GANs (cGANs), where the generator is restricted to modifying a fixed set of pre-defined attributes. To tackle this problem, they proposed an OpenGAN architecture. The unique selling point of this architecture is per input sample conditioning with feature maps taken from a metric space. OpenGAN seeks to generate images with attributes or attribute combinations that were not explicitly present in the original training data.

The metric learning model extract semantic information transfers to out of distribution novel classes. This allows generator to create samples outside of training distribution, thus leading to demonstrate that random sampling of metric space could produce high-quality samples. This method uses interpolation in feature space and latest space. This metric space has the property that similar images in terms of semantic content have feature representations that are closer together. Instead of the entire image being mapped to a single point in this feature space, it uses feature maps at a more granular (possibly per-pixel or per-object) level.

This gives fine control over which specific image elements are manipulated based on the attributes. It aims to achieve a clean separation between an image's core content and its attribute representations. This disentanglement is crucial for flexible image manipulation. If content and attributes are entangled in the generator's representation, changing one would likely introduce unwanted modifications to the other. This results in visually pleasing and semantically perfect transformations in image space.

Experiments shows that performance of classifiers can be greatly improved by augmenting training data with OpenGAN produced samples even when classes are outside of GAN training distribution. The architecture is very different from encoder-decoder GAN architectures, diverges from many classic GAN architectures that rely on the generator learning image-to-image translation with paired examples. It uses Hinge loss for adversarial component and MSE for feature loss component. One of the unique points is OpenGAN cleanly splits content information into two distinct spaces without need to prior information. OpenGAN also doesn't match feature space over entire dataset, but conditions on a per-feature basis. Thus, it allows testing data generation by conditioning on specific source image or allows random sampling from the metric feature space.

This brings us to end of discussion for recent GAN based Super Resolution methods.

2.5 A review of real-world applications using Generative Adversarial Network based Super Resolution methods

In this section, I will briefly discuss about GAN models which has found various practical uses in variety of fields. Most of these models deal with SR in one way or another.

Two Pathway Generative Adversarial Network (TPGAN) (Huang et al., 2017) has found its uses in photorealistic frontal view synthesis. It uses a combination of symmetry loss, adversarial loss and identity preserving loss. It has also found uses in downstream tasks such as face recognition and attribute estimation. (Bulat et al., 2018) proposed High-to-Low GAN model (HLGAN) has found uses in effectively increasing quality of real-world low-resolution images where input is not artificially generated low-resolution image. This model deal with real-world nuisance by using unpaired image data during training.

The GAN model proposed by (Indradi et al., 2019) has uses in generating face images and attributes to gain more facial information. This model uses inception residual network to stabilize

training and improve generated image quality. The AttentionalGAN model proposed by (Pathak et al., 2019) has uses in online website Expedia to generate SR images to display on website. This model uses a private dataset to train a GAN with attentional features from scratch results in better output image quality compared to SRGAN. This model also suggests a distributed algorithm to speed up GAN training by roughly a factor of five.

The EUSR-PCL GAN model suggested by (Cheon et al., 2019) uses content loss with discrete cosine transform coefficients loss plus differential content loss. The model has found uses in cases where high perceptual performance is required and slight distortion is acceptable. The AEGAN model proposed by (Guo et al., 2019) has found uses in cases where requirement is to directly generate high-resolution images using GAN. It involves an autoencoder for learning finer high-level structure of real images and use a denoiser to show photo-realistic details for generate SR image. It uses low-dimensional embedding to close the distribution gap between real data and input noise.

The JSIGAN model proposed by (Kim et al., 2019) found uses in to convert legacy low-resolution standard dynamic range (SDR) videos to high-resolution high dynamic range (HDR) videos in home television appliance space by performing fine detail restoration and contrast enhancement. This model involves adversarial loss with detail GAN loss to create a total loss function. The PGAN model proposed by (Mahapatra et al., 2019) found uses in medical image SR space, especially in more accurate detection of landmarks and pathologies. It involves triplet loss to perform stepwise image quality enhancements by using output of last stage as baseline to current stage.

The GAN model proposed by (Sood et al., 2019) found uses in medical Magnetic Resonance (MR) images SR space. It involves a combination of pixel wise MSE loss, VGG loss, content loss, adversarial loss and perceptual loss function. The KernelGAN model proposed by

(Bell-Kligler et al., 2019) found uses in cases where downscaling kernel is unknown. This model is fully unsupervised and doesn't require training data (other than input image itself) and is pluggable into existing SR methods.

The IPFSRGAN model proposed by (Li et al., 2020) found uses in cases where preserving identity of face images is important while reconstructing realistic ST face images from low-resolution images. This model pushes generator to learn and infer finer facial details. This model is useful in face verification tasks. The GAN model proposed by (Bhatia and Kumar, 2020) is useful in performing SR microscopy operations. It has tremendous uses in enhancing microscopic images. The paper presented by (Singh and Raza, 2020) shows uses of GAN in medical image generation using various GAN models such as DCGAN, LAPGAN, CycleGAN, Pix2Pix and UNIT framework.

The DSRGAN model proposed by (Demiray et al., 2020) shows uses of GAN in increasing resolution of a Digital Elevation Model (DEM) dataset which are used in road extraction, flood mapping, hydrological modelling and surface analysis. The model proposed by (Cao et al., 2020) shows uses in lossless image compression algorithms. This model uses entropy coding to create SR based compression module with SOTA compression rates. The Wavelet-SRGAN model proposed by (Yang et al., 2020) find uses in cases where requirement is to have SR images with richer high-frequency details such as medical field, remote sensing and criminal investigation. This model reconstructs and produces SR images with rich global as well as local texture details.

The AutoGAN model proposed by (Fu et al., 2020) has uses in devices with limited computing power, network bandwidth and memory space such as mobile devices. This model deals with compression of GAN model by performing end to end discovery for efficient generator within the target device computing envelope. AutoGAN is standalone, completely automatic and generally applicable to various GAN models.

This brings us to end of discussion for real-world application using GAN based Super Resolution methods.

2.6 Tabular representation of PSNR and SSIM metrics values used in CNN and GAN models

Figure 2.13 lists down PSNR and SSIM metrics values (if available) over common dataset used in literature for various CNN models discussed in this thesis so far.

CNN Models	Datasets											
4x Scaling	Set 5		Set 14		BSD100		Urban100		DIV2K		Manga109	
	I	II	I	II	I	II	I	II	I	II	I	II
SRCNN	30.48	0.8628	27.50	0.7513	26.90	0.7103	24.52	0.7226	29.33	0.8090	27.66	0.8580
IRCNN	-	-	-	-	-	-	-	-	-	-	-	-
ESPCN	-	-	-	-	-	-	-	-	-	-	-	-
REDNet	31.51	0.8869	27.86	0.7718	27.40	0.7290	-	-	-	-	-	-
LapSRN	31.54	0.8866	28.09	0.7694	27.32	0.7264	25.21	0.7553	29.88	0.8250	29.09	0.8900
SRDenseNet	32.02	0.8934	28.50	0.7782	27.53	0.7337	26.05	0.7819	-	-	-	-
DDBPN	32.47	0.8980	28.82	0.7860	27.72	0.7400	26.38	0.7946	-	-	30.91	0.9137
RCAN	32.63	0.9002	28.87	0.7889	27.77	0.7436	26.82	0.8087	30.77	0.8459	31.22	0.9173
SRRAM	32.13	0.8932	28.54	0.7800	27.56	0.7350	26.05	0.7834	-	-	-	-
ZSSR	31.13	0.8796	28.01	0.7651	27.12	0.7211	-	-	-	-	-	-
SRMDNF	31.96	0.8925	28.35	0.7787	27.49	0.7337	25.68	0.7731	30.01	0.8278	30.09	0.9024
VDSR	31.35	0.8838	28.02	0.7678	27.29	0.7252	25.18	0.7525	29.82	0.8240	28.82	0.8860

Figure 2.14
Showing PSNR and SSIM metric value of CNN models. (I) means PSNR and (II) means SSIM metric value. (-) means data is not available publicly. (Anwar et al., 2019)

Figure 2.14 lists down PSNR and SSIM metrics values (if available) over common dataset used in literature for various GAN models discussed in this thesis so far.

GAN Models	Datasets																			
	Scaling Factor	Set 5		Set 14		BSD100		Urban100		Manga109		LS3D-W		NTIRE 2018 Track 2		DIV2K		EYEPACS	
		I	II	I	II	I	II	I	II	I	II	I	II	I	II	I	II	I	II
TPGAN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SRPGAN	4x	22.68	0.8800	22.50	0.7860	23.91	0.7460	20.00	0.7690	21.00	0.8600	-	-	-	-	-	-	-	-
SRPGAN	8x	19.14	0.7490	19.10	0.6590	21.55	0.6190	17.68	0.6070	18.68	0.7900	-	-	-	-	-	-	-	-
SRGAN	4x	29.40	0.8472	26.02	0.7397	25.16	0.6688	-	-	-	-	-	-	-	-	-	-	-	-
HLGAN	-	-	-	-	-	-	-	-	-	-	-	19.90	-	-	-	-	-	-	-
SAGAN	4x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	44.90	0.8200
SAGAN	8x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	36.20	0.7700
SAGAN	16x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	31.80	0.7100
SAGAN	32x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	28.60	0.6700
CmGAN	4x	-	-	-	-	-	-	-	-	-	-	-	-	24.93	0.6900	-	-	-	-
ProGANSR	4x	-	-	28.94	-	27.79	-	26.89	-	-	-	-	-	-	-	30.81	-	-	-
ProGANSR	8x	-	-	25.29	-	24.99	-	23.04	-	-	-	-	-	-	-	27.96	-	-	-
ESRGAN	4x	32.73	0.9011	28.99	0.7917	27.85	0.7455	27.03	0.8153	31.66	0.9196	-	-	-	-	-	-	-	-
KPSNRLGAN	4x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	26.81	0.7265	-	-
PGAN	4x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	46.10	0.9100
PGAN	8x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	40.30	0.8100
PGAN	16x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	36.80	0.7700
PGAN	32x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	31.90	0.7200
IESRGAN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AHGAN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
HUSR-PCLGAN	4x	31.57	0.8743	28.34	0.7967	27.11	0.7643	-	-	-	-	-	-	-	-	-	-	-	-
ATTENTIONALGAN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ICESRGAN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
AUDSGAN	4x	30.44	-	27.28	-	26.23	-	24.74	-	-	-	-	-	-	-	-	-	-	-
WAVELET-SRGAN	4x	30.63	0.8970	27.04	0.7700	26.91	0.7280	24.12	0.7490	-	-	-	-	-	-	-	-	-	-
DSRGAN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
OPENGAN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LSRGAN	4x	30.32	0.8560	26.46	0.7340	25.52	0.6800	24.73	0.7570	-	-	-	-	-	-	-	-	-	-
SPSRGAN	4x	30.40	0.8627	26.64	0.7930	25.51	0.6576	24.80	0.9481	-	-	-	-	-	-	-	-	-	-
SPSAGAN	4x	-	-	-	-	25.16	0.6344	-	-	-	-	-	-	-	-	-	-	-	-
ESRGAN+	4x	-	-	-	-	-	-	29.28	-	-	-	-	-	-	-	-	-	-	-
FCSRGAN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
IPTSRGAN	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Figure2.15
Showing PSNR and SSIM metric value of GAN models. (I) means PSNR and (II) means SSIM metric value. (-) means data is not available publicly or dataset is private.

2.7 Discussion

Exploration in super resolution space is still going strong with computer vision community. GAN models have introduced an exciting new dimension to SR image space. As part of literature review, I have identified various gaps in some of the current implementations. These gaps lead to open questions which should be further explored.

- I have noticed that Deep learning CNN architectures used in various SR image space doesn't bode well with deeper architecture. For example: SRCNN. This requires investigations and more testing to better understand gradients and training dynamics in deep architectures.
- Most of the existing GAN models doesn't provide visually pleasing perceptual SR image with high up-sampling factor of 8x and more. This should be further explored and tuned to produce great looking SR output with high up-sampling factors of 8x and beyond.

- The inherent nature of GAN models gives great flexibility to further improve the performance by exploring various combination of loss functions, filters, position of layers and training strategies. Most of the past research haven't explored these parameters in a profound way.
- GAN models have fundamental design decision to synchronize discriminator with generator. Just a handful of research have been performed so far to improve efficiency between generator and discriminator to improve training time and find better distributions for training sample.
- Most of the existing GAN models work on single space i.e. treat every feature a separate input and don't try to learn relationship between input. This could be further explored to find inherent correlation input. This could prove very useful in face attribute learning tasks.
- Loss functions used in GAN models have been mostly focused on image super resolution, face attributes and object identification tasks. It should be further explored in the area of image colorization and image semantic segmentation.
- Most of the existing GAN models are not mobile compute friendly and lacks compression to reduce overall compute footprint. It should be further explored by devising an application and device specific loss function to make a generalized GAN compression model. This will help push GAN SR in real time high framerate video streaming, conversion and edge computing.
- Most of the existing GAN models are in supervised learning space. GAN models in unsupervised space are few and not been explored so much. Unsupervised space should be further explored to create better GAN models.
- GAN based SR models tends to fail when there is extreme blur, pose distortion and occlusion. Recent work tries to improve on these but the success is at limited level. It should

be explored further to see if output image can be improved by understanding relationship between various input image to perceive and generate a relative output image.

- Most of the existing GAN models are using standard 3x3 or 5x5 spatial convolution filters. Research shows that using larger spatial filters increases computational cost. It should be explored further by using asymmetric convolution with better network architectures and training parameters.
- Most of the existing GAN models have used just one architecture for training. It would be interesting to train on ensemble of models where each model is using a different feature set, and finally aggregate all models to get final SR image output.
- GAN model generally doesn't insert noise in input image. Recent research has shown that adding noise can result in better perceptual quality of generated output. It should be further explored by trying various noise types and variations.
- GAN models have mostly dealt with 2D face priors. It would be interesting to work on 3d face priors.
- In case of cross domain image to image training and translation, GAN models don't show promising result and show unable to retain important information. Cross domain training should be further explored.
- A better encoder design that can represent natural high-resolution or low-resolution image distribution in latent space could produce better SR output. This should be further explored.

2.8 Summary

We have come to end of literature review. In this literature review, I have started with brief overview of Super Resolution, why it is still a challenging problem in computer vision, what is GAN, why GAN has taken centre stage for SR image space and finally discussing feature extraction and its usefulness in SR space. Next, I have discussed various SR methods available in existing literature, their network architecture, loss function used and their advantage or disadvantage. This provided us a very good overview and insight into SR history and how it has come so far. I have kept GAN based SR methods out of this section because it is discussed in details at next section. Further expanding the SR review, I went deep into some of most profound DCNN based SR methods. These methods have paved the way for recent and future work in SR space.

Building on SR knowledge, I have discussed GAN model in SR space. I have written in length about nine profound and most impactful GAN models. In includes problem space, architecture, loss functions and their performance compared to SOTA work. Further, I briefly touched on more than fifteen recent GAN models which have outperformed SOTA work so far. Next, I have discussed various real-world applications which have used GAN based SR methods to provide real value in day to day used computing needs, such as medical imaging, image compression, real-time video streaming, edge computing, face attributes learning, video surveillance and terrain mapping to list a few.

Finally, I have tabulated PSNR and SSIM metrics values of discussed SR models for both CNN and GAN based methods (wherever available). The data is collected from publicly available literature and research papers. This makes it easy for reader to see performance of SR methods at once glace. I have finished literature review by discussing gaps and challenges found in existing GAN based SR methods. I also list down future direction which should be explored to generate even better-looking perceptual output with low computing footprint.

CHAPTER III:

RESEARCH METHODOLOGY

3.1 Introduction

The purpose of proposed model is to hopefully improve state-of-the-art in generative adversarial network models for super resolution on feature extraction by performing various experiments on large challenging real-life face images datasets. Proposed technical implementation is focused on performing various experiments on Feature Extractor and hyper parameters (filters, loss functions and upscaling factors) on a base GAN model.

The proposed model performs experiments by using pre-trained transfer learning backbones on large challenging real-life face image datasets. This thesis uses SRGAN (Ledig et al., 2017) as a base GAN model to perform experiments on feature extraction. The hypothesis is about substituting VGG19 by VGGFace2 and EfficientNet-B0 in SRGAN. Applying this hypothesis on real-life human face dataset should improve quantitative metrics values and should improve observable visual quality of generated images for feature extraction. The conceptual architecture of SRGAN model used in thesis is listed in figure 3.1.

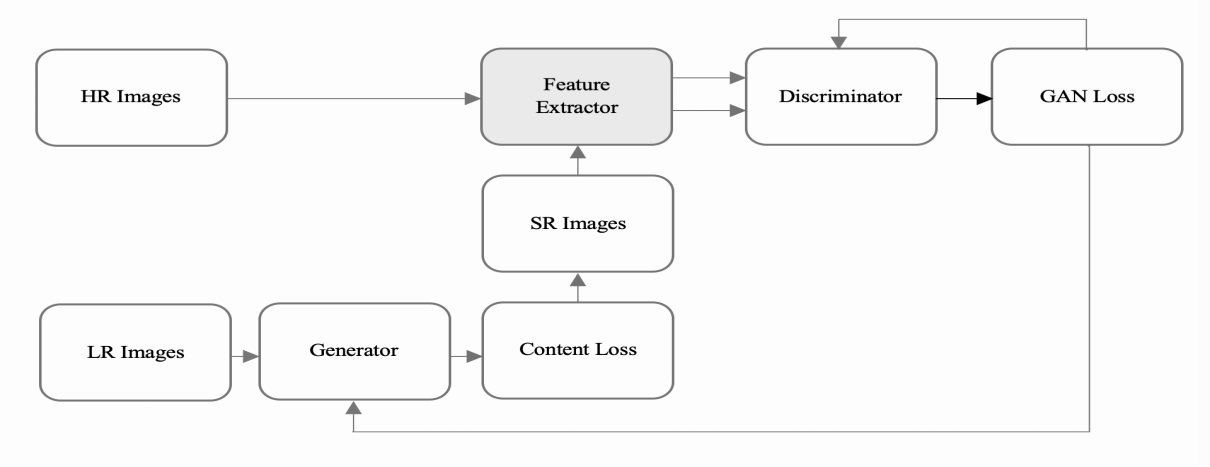


Figure3.1
Conceptual SRGAN model architecture.

The methodology for this research is divided in to five steps:

- Dataset understanding and preparation
- Proposed model
- Train and test
- Model evaluation
- Result analysis and comparison

3.2 Dataset understanding and preparation

In this proposed model, three datasets of large challenging face images are used. Datasets contain large number of face images having large variation faces, pose, age, illumination, profession and ethnicity. The test data in datasets are partitioned into training, validation and testing sets. The datasets are CelebA (Liu et al., 2018), LFW (Lfw, 2013) and Simpson (alexattia, 2018).

- CelebA – The complete name is CelebFaces Attributes (CelebA) Dataset. It is a collection of face images with large pose variations, background clutter, diverse people, supported by a large quantity of images and rich annotations. The dataset contains over 200 thousand number of face images of various celebrities, more than 10 thousand unique identities, 5 landmark locations and 40 binary attribute (1 represent positive and -1 represent negative) annotations per image. The dataset has images partitioned for training, validation and testing sets. It also has bounding box for each image to represent width and height.
- LFW – The complete name is Labeled Faces in the Wild Dataset. It is collection of face images designed for studying the challenges of unconstrained face recognition. It contained over 13 thousand images of more than 5 thousand people. More than 1500 people have two or more picture in dataset. The deep-funneled version produces superior results. Every image is 250x250 jpg, centered and detected using the OpenCV implementation of Viola-

Jones face detector. It also provides names of each face in dataset and number of images each face has. It has more than one recommended set for training and testing (i.e. pairs versus people).

- Simpson – The complete name is Simpson Characters Dataset. This is face image dataset having images from popular cartoon character series Simpson. There are total of 20 characters with each having 400-2000 pictures. Each picture also has bounding boxes for each character.

This proposed model uses image aligned versions of CelebA, LFW and Simpson. This almost eliminates dataset preparation activities and speed up technical implementation work. The image aligned version of datasets comes pre-partitioned into training, validation and testing sets. Overall, use of three different and challenging face datasets gives a very good opportunity to our proposed model to also experiments with cross domain learning. This is due to reason that two of our face datasets are not real-life human faces. By running our model on these datasets, it will improve model performance in diverse real-life use cases with face images.

3.3 Proposed model

The proposal is to perform experiments on Feature Extractor and hyper parameters (filters, loss functions and upscaling factors) on a base SRGAN model with pre-trained backbones to increase learning speed. Generator and discriminator are basic blocks of GANs (Goodfellow et al., 2014). The proposed model uses Keras (Keras.io, 2020), TensorFlow (TensorFlow, n.d.), OpenCV (OpenCV, n.d.), VGG19, VGGFace2 and EfficientNet. In the proposed model, these experiments are conducted on base architecture listed in figure 3.2, such as:

- SRGAN with VGG19 (Feature Extractor).

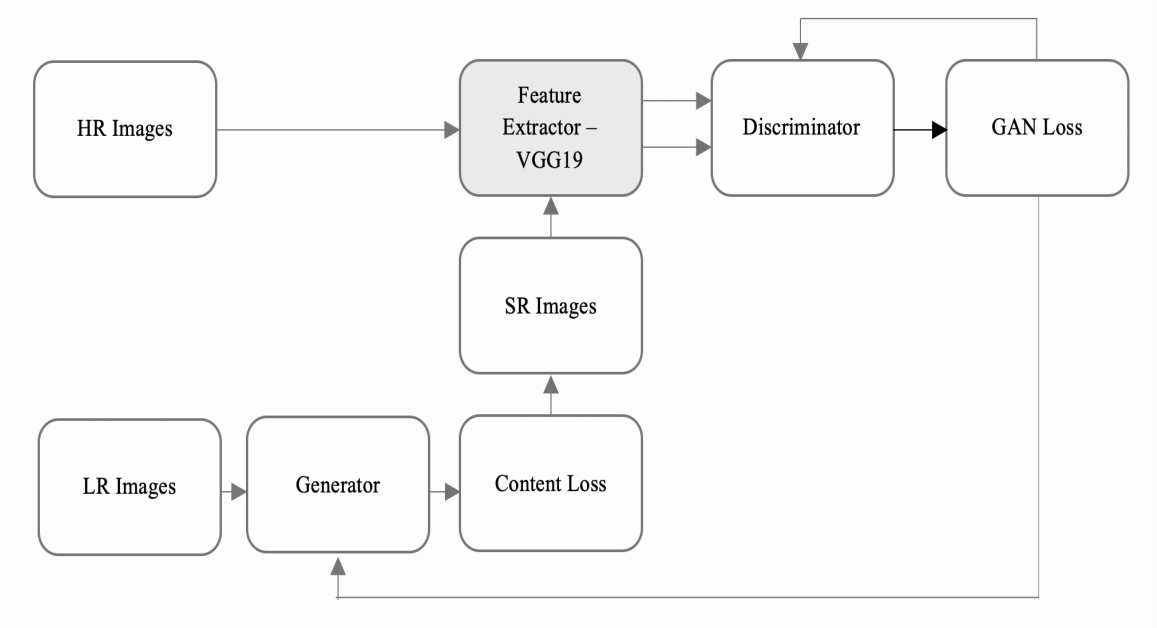


Figure3.2
Proposed SRGAN model with VGG19 feature extractor.

- SRGAN with VGGFace2 (Feature Extractor) by substituting VGG19.

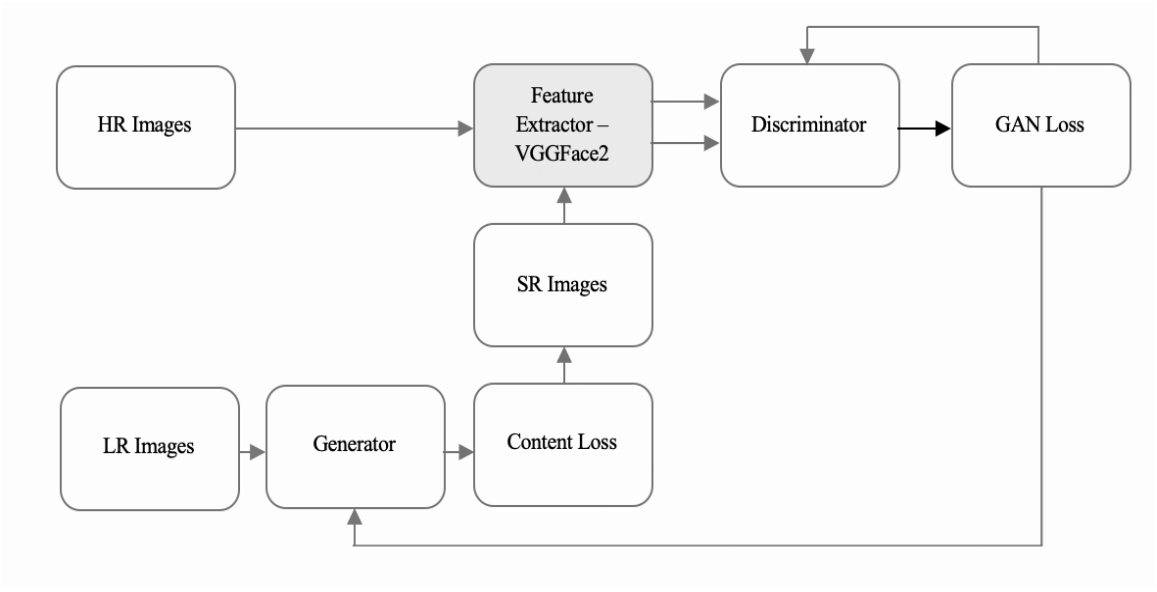


Figure3.3
Proposed SRGAN model with VGGFace2 feature extractor.

- SRGAN with EfficientNet-B0 (Feature Extractor) by substituting VGG19.

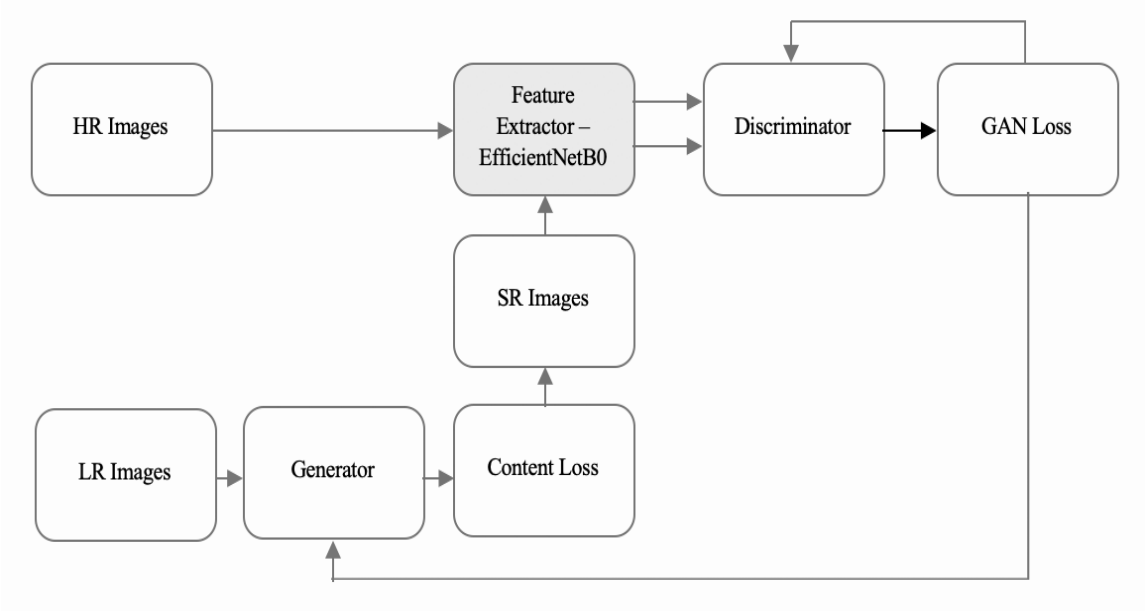


Figure3.4
Proposed SRGAN model with EfficientNet-B0 feature extractor.

- Experiments on hyper parameters such as content loss, GAN loss and filters.

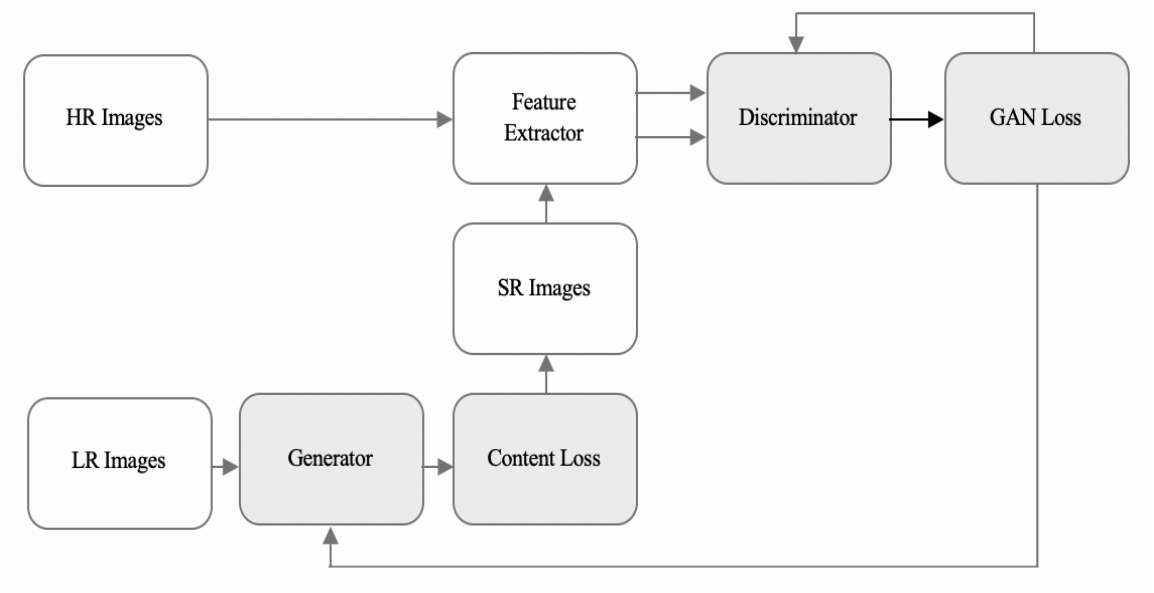


Figure3.5
Proposed SRGAN model with experiments on hyper parameters.

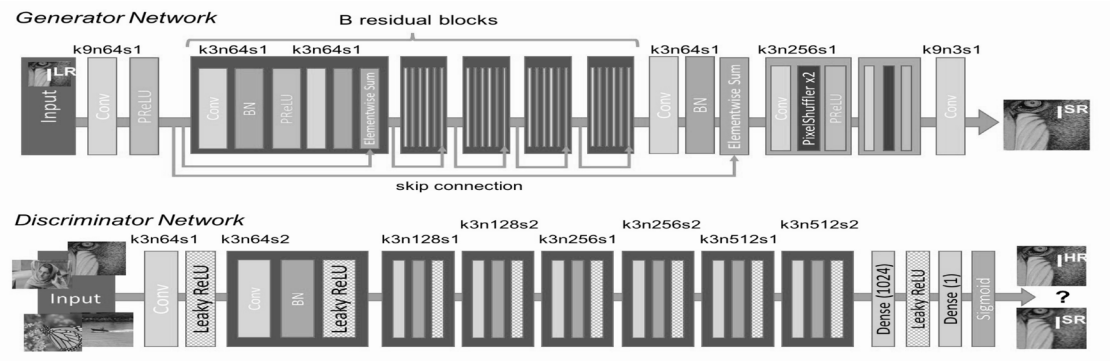


Figure3.6
SRGAN network (Ledig et al., 2017) showing filters on which various experiments are proposed to be performed.

3.3.1 Loss functions

The proposed model involves six types of loss functions and combine them to represent Generator-Discriminator loss in result evaluation and interpretation.

- MSE loss – It works if the goal is to generate a SR picture having the best pixel colours conformity with the ground truth picture. However, in real-life scenarios it might be necessary to concentrate on the structure or relief of the picture.

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} |I(i,j) - K(i,j)|^2$$

Eq 3.1 MSE loss equation

- Perceptual loss – It is a weighted sum of the content loss (l^{SR}_X) and adversarial loss (l^{SR}_{GEN}).

$$l^{SR} = l^{SR}_X + 10^{-3} l^{SR}_{Gen}$$

Eq 3.2 Perceptual loss equation

- Content loss – It is a pixel-wise loss between each pixel in real image and a pixel in generated image.

$$l^{SR}_{MSE} = \frac{1}{r^2 WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} (I^{HR}_{x,y} - G_{\theta_G}(I^{LR})_{x,y})^2$$

Eq 3.3 Content loss equation

Here, 'r' is down sampling factor, 'W' is tensor width of low-resolution image, 'H' is tensor height of low-resolution image, 'x-y' represents pixel coordinates, 'G' is generating function, 'I^{HR}' is high-resolution image, 'I^{LR}' is low-resolution image and '$G_{\theta G}$' represents function parameterized with θ_G.

- VGG loss – It is the Euclidean distance between the feature maps of the generated image and the real image. Here, '$\phi_{i,j}$' represents the feature map obtained by the j^{th} convolution (after activation) before the i^{th} max-pooling layer within the VGG19 network.

$$l^{SR}_{VGG/i,j} = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{LR}))_{x,y})^2$$

Eq 3.4 VGG loss equation

- Adversarial loss – It is calculated based on probabilities provided by discriminator.

$$l_{Gen}^{SR} = \sum_{n=1}^{N} -\log D_{\theta_D}(G_{\theta_G}(I^{LR}))$$

Eq 3.5 Adversarial loss equation

Here, 'D' is discriminator function, '$D_{\theta D}$' represents function parameterized with θ_D.

- Generative loss – This is present in most GAN model because of notion that generator tries to minimize while discriminator tries to maximize.

3.3.2 Upscaling factor

The proposed model aims to start with 4x upscaling. Based on experiment results, higher upscaling (8x or more) may be attempted.

3.4 Train and Test

The proposed model uses image aligned versions of CelebA, LFW and Simpson datasets. The image aligned version of datasets comes pre-partitioned into training, validation and testing sets. Model training will use epoch count of greater than 200 in initial runs and will be increased in subsequent runs. Model training uses training image set of datasets while model testing uses testing set of datasets. The batch size is initially kept at 8. Optimization is managed by Adam optimizer. By running our model on these datasets, it will improve model performance in diverse real-life use cases with face images.

3.5 Model evaluation

The model evaluation focuses on Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM) metrics values to evaluate performance of proposed model. These are very popular metrics (Horé and Ziou, 2010) used in all super-resolution GAN models.

- PSNR is defined as ratio between maximum possible power of signal and power of corrupting noise. A high value of PSNR means more noise removed. It is little biased towards over smooth (i.e. blurry) results.

$$PSNR = 10\log_{10}\left(\frac{MAX_I^2}{MSE}\right)$$

Eq 3.6 PSNR equation

Here, 'MAX_I' represents maximum possible power of a signal of image 'I' and 'MSE' represents mean squared error pixel by pixel.

- SSIM is defined as measurement of perceptual difference between two similar images. It considers similarity of edges i.e. high frequency content between denoised and ideal image. It is a better metric compared to PSNR but more complicated to compute.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

Eq 3.7 SSIM equation

Here, 'μ_x' represents average value for the first image, 'μ_y' represents average value for the second image, 'σ_x' represents standard deviation for the first image, 'σ_y' represents standard deviation for the second image and 'c_1,c_2' are coefficients.

3.6 Result analysis and comparison

Result analysis and comparison focuses on graphs plots showing PSNR and SSIM metric values. These quantitative values will also be presented in a tabular format.

3.7 Summary

The proposed model tries to improve state-of-the-art in generative adversarial network models for super resolution on feature extraction by performing various experiments on large challenging real-life face images datasets. Technical implementation is focused on performing various experiments on Feature Extractor and hyper parameters (filters, loss functions and upscaling factors) on a base GAN model with pre-trained backbones to increase learning speed. The hypothesis is about substituting VGG19 by VGGFace2 and EfficientNet-B0 in SRGAN.

The proposed model involves six types of loss functions. These are MSE loss, perceptual loss, content loss, VGG loss, adversarial loss and generative loss. The proposed model suggests to perform various experiments on these loss functions.

This thesis uses image aligned versions of CelebA, LFW and Simpson. This almost eliminates dataset preparation activities and speed up technical implementation work. Overall, use of three different and challenging face datasets gives a very good opportunity to our proposed model to also experiments with cross domain learning. This is due to reason that two of our face datasets are not real-life human faces. By running our model on these datasets, it will improve

model performance in diverse real-life use cases with face images. The model evaluation focuses on Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM) metrics values to evaluate performance of proposed model. The proposed model use of real-life human face dataset should improve quantitative metrics values and should improve observable visual quality of generated SR images.

CHAPTER IV:

ANALYSIS

4.1 Introduction

The proposed model uses three datasets of large challenging face images. Datasets contain large number of face images having large variation faces, pose, age, illumination, profession and ethnicity. These datasets are image aligned versions of CelebA (Liu et al., 2018), LFW (Lfw, 2013) and Simpson (alexattia, 2018) datasets. The image aligned version of datasets comes pre-partitioned into training, validation and testing sets. This eliminates dataset preparation activities and speed up technical implementation work. Use of three different and challenging face datasets gives a very good opportunity to experiments with cross domain learning and improve model performance in real world. The careful selection of CelebA, LFW, and Simpson datasets reveals a strategic approach for training robust models tailored for feature extraction. Each dataset possesses valuable characteristics that contribute to different aspects of training and address potential real-world challenges. A brief overview of datasets is listed below.

- CelebA – The complete name is 'CelebFaces Attributes' dataset. It is a collection of face images with large pose variations, background clutter, diverse people, supported by a large quantity of images and rich annotations. CelebA's strengths lie in its extensive size and rich annotations. The dataset's large-scale nature fosters model generalization, as the model is exposed to a vast array of facial appearances under diverse conditions. The annotations, including attributes like pose, age, and background information, can facilitate several strategies.

  The annotations can be used to guide the discriminator network within the GAN architecture, encouraging it to discern subtle differences in super-resolved images beyond pixel-level fidelity. This could lead to better preservation of identity-specific and pose-

related features during the up-sampling process. CelebA's annotations make it suitable for training conditional GAN variations (cGANs). Conditioned on specific attributes, the model could learn to specialize in generating super-resolved images with desired characteristics (e.g., preserving a specific pose or age bracket).

- LFW – The complete name is ‘Labeled Faces in the Wild’ dataset. It is collection of face images designed for studying the challenges of unconstrained face recognition. It is known for its "in the wild" nature, presents significant challenges in unconstrained face recognition. By training models on LFW, there are several implications for feature extraction. Super-resolution aims to reconstruct a high-resolution (HR) image from a low-resolution (LR) input.

  LFW's unconstrained nature simulates real-world camera limitations, forcing the model to learn how to recover detail loss due to factors like blur, compression artifacts, and noise. This directly enhances the model's ability to extract meaningful features even from degraded input. Training on LFW alongside datasets like CelebA promotes transfer learning. The model must learn to reconcile different image distributions, making its extracted features more robust to variations encountered in real-world applications.

- Simpson – The complete name is ‘Simpson Characters’ dataset. This is Face image dataset having images from popular cartoon character series Simpson. This dataset introduces intriguing possibilities for domain adaptation and potential bias mitigation. The cartoon-like nature of Simpson creates a significant domain gap with CelebA and LFW. Training on this dataset, forces the model to discover the essential 'structural' components of faces.

  This knowledge can be leveraged through domain adaptation techniques to improve super-resolution on other stylized data (e.g., paintings, sketches). CelebA and LFW primarily contain real human faces and might carry inherent biases related to ethnicity,

gender representation, etc. Simpson can serve as a counterbalance. Its simplified visual style could help the SRGAN learn a more generalized representation of facial features, potentially reducing biases when applied to real-world data.

In this chapter, features of each dataset are presented in detail and analyzed in correlation with model objectives.

4.2 Dataset description

In this section, features of each dataset are presented in details. Let's start with CelebA dataset, followed by LFW and finally, Simpson dataset.

4.2.1 CelebA dataset

CelebA dataset is a collection of face images with large pose variations, background clutter, diverse people, supported by very large quantity of images and rich annotations. This dataset supports proposed model objectives very well by providing a complex and ginormous collection of facial images to support feature extraction and image generation. The sheer scale and quantity of facial images allow opportunities to tinker with model parameters.

Figure4.1
CelebA dataset (Liu et al., 2018) sample attributes.

The image aligned version of CelebA (Jessica Li, 2018a) dataset contains over 200 thousand number of face images of various celebrities, more than 10 thousand unique identities, 5 landmark locations and 40 binary attribute annotations per image. The dataset has images partitioned for training, validation and testing sets. It also has bounding box for each image to represent width and height. The properties of the CelebA dataset make it a particularly valuable resource for training SRGAN models aimed at robust feature extraction.

A large dataset significantly reduces the risk of overfitting to a limited set of facial features. It forces the model to learn a broader representation, enhancing its capacity to extract meaningful features even from previously unseen faces. Diverse identities expose models to the nuanced differences in facial structure, expressions, and textures. This encourages the model to develop fine-grained feature representations essential for tasks like identity preservation during super-

resolution. Landmarks act as anchors, directing the model's attention to key facial components (e.g., eyes, nose, mouth).

This promotes the preservation of their relative positions and proportions in the super-resolved image, improving the geometric accuracy of facial features. The bounding boxes allow the model to focus on the region containing the face. This can increase computational efficiency and prevent the model from being distracted by irrelevant background information. Additionally, bounding boxes can be used to isolate specific facial regions if desired, facilitating localized feature extraction. The 40 binary attributes offer a rich source of high-level semantic information. The model can be trained to associate these attributes with specific patterns in the super-resolved images.

This fosters the extraction of semantically meaningful features beyond raw pixel data. Attribute annotations enable the development of conditional or attribute-aware models. The model could be trained to generate super-resolved images that enhance or modify specific attributes, opening up possibilities for controlled image manipulation and analysis. The dataset's preparation into training, validation, and testing sets offers several practical advantages. Pre-defined splits streamline model development and benchmarking. They ensure comparability of results when evaluating different model's architectures or training procedures. The validation set is instrumental in hyperparameter optimization, allowing researchers to fine-tune their models learning process without compromising the integrity of the final test set evaluation. The details of CelebA dataset are given in table 4.1, along with features of dataset.

Feature	Details
Dataset size	Approximate 1GB

Unique attributes of dataset	Large pose variations Background clutter Foreground variations Diverse racial features Diverse ethnicity features Hair styles and types Facial features Facial expressions Diverse age groups Lighting variations
Number of face images	202599
Image demography	Celebrity images
Number of unique identities	10177
Names of unique identities	Hidden, not provided
Number of attribute annotations per image	40
Attribute annotations type	Binary (1 represents positive and -1 represents negative)
Name of attributes	5_o_Clock_Shadow Arched_Eyebrows Attractive Bags_Under_Eyes Bald Bangs

	Big_Lips Big_Nose Black_Hair Blond_Hair Blurry Brown_Hair Bushy_Eyebrows Chubby Double_Chin Eyeglasses Goatee Gray-haired Heavy_Makeup High_Cheekbones Male Mouth_Slightly_Open Mustache Narrow_Eyes No_Beard Oval_Face Pale_Skin Pointy_Nose Receding_Hairline

	Rosy_Cheeks Sideburns Smiling Straight_Hair Wavy_Hair Wearing_Earrings Wearing_Hat Wearing_Lipstick Wearing_Necklace Wearing_Necktie Young
Number of landmark locations	5
Landmark location types	Left eye Right eye Nose Left mouth Right mouth
Image alignments	Cropped and aligned to face
Image bounding box	X1-Y1 represent upper left point coordinate of bounding box. Width and height represent width and height of bounding box.
Number of data partitions	3

Data partition types	Training Validation Testing
Image distribution in data partitions	Training distribution: Images 1-162770 Validation distribution: Images 162771-182637 Testing distribution: Images 182638-202599
Number of data files	5
Data files name	Imgalignceleba.zip: It represents all images. Listevalpartition.csv: It represents recommended data partition. Listbboxceleba.csv: It represents bounding box information. Listlandmarksalign_celeba.csv: It represents landmarks and landmark coordinates. Listattrceleba.csv: It represents attribute labels for each image.
Number of columns	59
Number of data types	2
Data type name and column count	Integer: 55 String: 4

Table 4.1 Features and details of CelebA dataset

The image aligned CelebA dataset comes pre-partitioned into training, validation and testing sets. This eliminates dataset preparation activities such as variable elimination, handling of

categorical variables, missing value identification and treatment. This speeds up technical implementation work.

4.2.2 LFW dataset

LFW dataset is collection of face images designed for studying the challenges of unconstrained face recognition. It is widely recognized and used in benchmarks for face verification and feature extraction. The image aligned deep-funneled version of LFW (Jessica Li, 2018b) dataset contains over 13 thousand face images of more than 5 thousand people. More than 1500 people have two or more picture in dataset. While smaller than CelebA, LFW's 13,000+ images and 5,000+ identities still represent a substantial dataset. The presence of individuals with multiple images is particularly valuable for training models tasked with extracting identity-preserving features.

Figure4.2
Original LFW dataset (Lfw, 2013) pictorial illustration.

The deep-funneled version produces superior results. Every image is centered and detected using the OpenCV implementation of Viola-Jones face detector. The detector returns a cropping region automatically enlarged by a factor of 2.2 in all dimension to capture more of head. Finally, it scaled to a uniform size. It also provides name of each face in dataset and number of images each face has. It has multiple recommended set for training and testing (i.e. pairs versus people). This

dataset supports proposed model objectives very well by providing a complex and sizable collection of facial images to support feature extraction and image generation.

LFW is designed for unconstrained face recognition underscores the variability and complexity inherent in the dataset. Unlike controlled datasets, LFW includes faces with diverse head orientations, occlusions, and natural expressions. This forces the model to extract pose-invariant and expression-robust features to maintain a consistent representation of identity while super-resolving the image. Images in LFW exhibit substantial environmental variations. The model must learn to differentiate between core facial features and artifacts caused by lighting, shadows, and complex backgrounds.

"In the wild" images often suffer from blur, compression artifacts, low resolution, and noise. Training on LFW enhances the SRGAN's ability to reconstruct high-quality features even from significantly degraded input. The multi-stage alignment and scaling process in deep-funneled LFW dataset emphasizes facial details by centering and standardizing images. This allows the model to concentrate on learning intrinsic facial features rather than being overwhelmed by extraneous variations.

Deep-funneling introduces a level of quality control and normalization to the LFW dataset. This can aid in reducing batch-to-batch variability during model training, potentially leading to improved stability and convergence. The details of LFW dataset are given in table 4.2, along with features of dataset.

Feature	Details
Dataset size	Approximate 150MB
Unique attributes of dataset	Large pose variations

	Background clutter Foreground variations Diverse racial features Diverse ethnicity features Hair styles and types Facial features Facial expressions Diverse age groups Lighting variations
Number of images	13233 Centered on Face
Image demography	Famous people images
Number of unique values	5749
Number of faces with more than one photo	1680
Names of unique identities	Not hidden
Image size	250 x 250 Pixel
Number of attributes	73
Name of attributes	Person Imagenum Male Asian White

	Black Baby Child Youth Middle Aged Senior Black Hair Blond Hair Brown Hair Bald No Eyewear Eyeglasses Sunglasses Mustache Smiling Frowning Chubby Blurry Harsh Lighting Flash Soft Lighting Outdoor Curly Hair

	Wavy Hair Straight Hair Receding Hairline Bangs Sideburns Fully Visible Forehead Partially Visible Forehead Obstructed Forehead Bushy Eyebrows Arched Eyebrows Narrow Eyes Eyes Open Big Nose Pointy Nose Big Lips Mouth Closed Mouth Slightly Open Mouth Wide Open Teeth Not Visible No Beard Goatee Round Jaw Double Chin Wearing Hat

	Oval Face Square Face Round Face Color Photo Posed Photo Attractive Man Attractive Woman Indian Gray Hair Bags Under Eyes Heavy Makeup Rosy Cheeks Shiny Skin Pale Skin 5 o' Clock Shadow Strong Nose-Mouth Lines Wearing Lipstick Flushed Face High Cheekbones Brown Eyes Wearing Earrings Wearing Necktie Wearing Necklace
Image alignments	Cropped and aligned to center

Number of data partitions	3
Data partition types	Training Validation Testing
Number of data files	11
Data files name	lfw-deepfunneled.zip: It represents file containing images. lfwallnames.csv: It represents name of each face and number of faces. lfwreadme.csv: It represents readme file with column metadata information from original LFW dataset. pairs.csv: It represents randomly created splits for 10-fold cross validation for pairs. people.csv: It represents randomly created 10-fold cross validation for faces. matchpairsDevTest.csv: It represent 500 testing sets for pairs. matchpairsDevTrain.csv: It represents 100 training sets for pairs. mismatchpairsDevTest.csv: It represents 500 test sets for face pairs. mismatchpairsDevTrain.csv: It represents 100 training sets for face pairs.

	peopleDevTest.csv: It represents 1711 people and 3708 images for people test set. peopleDevTrain.csv: It represents 4038 people and 9525 images for people train set.
Number of columns	136
Number of data types	2
Data type name and column count	Integer: 55 String: 4

Table 4.2 Features and details of LFW dataset

The image aligned deep-funneled LFW dataset comes pre-partitioned into multiple training, validation and testing sets. This eliminates dataset preparation activities such as variable elimination, handling of categorical variables, missing value identification and treatment This speeds up technical implementation work.

4.2.3 Simpson dataset

Simpson dataset (alexattia, 2018) is collection of face images having images from popular cartoon character series Simpson. There are total of 20 characters with each having 400-2000 pictures. Each picture also has bounding boxes for each character. The inclusion of the Simpson dataset introduces intriguing dimensions to this research, complementing the benefits of CelebA and LFW.

Figure4.3
Simpson dataset (alexattia, 2018) sample cartoon images.

This dataset supports proposed model objectives very well by providing a complex and huge collection of cartoon facial images to support feature extraction and image generation. The large scale and quantity of cartoon facial images allow opportunities to tinker with model parameters. This also improves model overall image generation and feature extraction capabilities in real world scenario.

The starkly different visual style of cartoon faces in Simpson, compared to real-world faces in CelebA and LFW, creates a significant domain gap. Training SRGAN models to successfully super-resolve Simpson faces pushes the model to learn the core structural elements of a face. This

encourages the discovery of features that transcend stylistic variations, potentially making the model more robust when applied to diverse image types. Techniques like domain adaptation and transfer learning are explored to leverage knowledge gained from this dataset to improve model performance on real-world image domains.

CelebA and LFW, containing primarily real-world human faces, may carry inherent biases based on factors like ethnicity, gender representation, and age distribution. The simplified and stylized visual nature of the Simpson dataset might help counteract these biases. Training on Simpson introduces a more abstract representation of facial features into the model. This could potentially reduce the model's over-reliance on cues specific to particular ethnicities or demographic groups present in real-world datasets.

This dataset is used as testbed for systematic bias analysis. By controlling aspects like character design, this could isolate the impact of specific visual features (e.g., skin tone, facial proportions) on SRGAN's models feature extraction and image generation processes. The details of Simpson dataset are given in table 4.3, along with features of dataset.

Feature	Details
Dataset size	Approximate 1GB
Unique attributes of dataset	Large pose variations Background clutter Foreground variations Diverse facial features Hair styles and types Facial expressions

	Diverse facial proportions Lighting variations Cartoon characters Multiple faces in one image
Number of face images	41909
Image demography	Cartoon images
Number of unique identities	43
Names of unique identities	Not hidden
Number of attribute annotations per image	6
Name of attributes	Character name File path Bounding box coordinates: X1-Y1 and X2-Y2
Image alignments	Cropped and aligned to face
Image bounding box	X1-Y1 represent upper left corner coordinate of bounding box. X2-Y2 represent lower right corner of bounding box.
Number of data partitions	2
Data partition types	Training Testing
Number of data files	4
Data files name	Simpson_dataset.tar: It represents all images. Kaggle_simpson_testset: It represents preview of image dataset.

	Annotations.txt: It represents bounding box information for each character. Weights.best.hdf5: It represents compounded weights for easier prediction in kernels.

Table 4.3 Features and details of Simpson dataset

The image aligned Simpson dataset comes pre-labeled and error free. This eliminates dataset preparation activities such as variable elimination, handling of categorical variables, missing value identification and treatment. Reduced dataset preparation overhead frees up time to focus on core model architecture design, loss functions, and hyperparameter optimization. The significant number of images per character allows for extensive experimentation. This allows to further experiments and investigate the impact of factors like training set size, varying character proportions within the dataset, and class imbalance on the model's feature extraction capabilities. This speeds up technical implementation work. Finally, starting with a clean and well-structured dataset lowers the risk of errors or inconsistencies.

4.3 Summary

In this chapter, three datasets of large challenging face images are presented and their features are analyzed. Use of three different and challenging face datasets gives a very good opportunity to experiments with cross domain learning and improve model performance in real world. These datasets support proposed model objectives very well by providing a complex and huge collection of facial images to support feature extraction and image generation. The large scale, quantity and variations of facial images allow opportunities to tinker with model parameters. This

also improves model overall image generation and feature extraction capabilities in real world scenario.

These datasets contain large number of human and cartoon face images having large variations in illumination, diverse profession, large pose variations, background clutter, foreground variations, diverse racial features, diverse ethnicity features, hair styles and hair types, facial expression, diverse facial features and facial proportions, multiple faces in one image over diverse age groups and lighting variations. These datasets are image aligned versions of CelebA dataset, image aligned deep-funneled version of LFW dataset and image aligned version of Simpson dataset.

The image aligned version of datasets comes pre-labeled and pre-partitioned into training, validation and testing sets. This eliminates dataset preparation activities such as variable elimination, handling of categorical variables, missing value identification and treatment. This speeds up technical implementation work.

The strategic selection of the CelebA, LFW, and Simpson datasets lays a powerful foundation for this research, with a strong emphasis on robust feature extraction and real-world adaptability. The datasets introduce a remarkable degree of diversity across multiple dimensions.

The combination of human faces (CelebA, LFW) with cartoon representations (Simpson) exposes models to a vast range of facial structures, textures, styles, and ethnicities. It fosters feature representations that capture both the commonalities of faces and the nuances critical for discrimination. Variations in pose, lighting, background clutter, expression, and even occlusions (LFW) force the SRGAN models to extract essential facial information even under challenging conditions. This translates directly to improved performance in real-world applications where images are rarely pristine.

The inclusion of both human and cartoon faces provides a unique opportunity for cross-domain learning and domain adaptation experimentation. This allows experimentations such as does training on one domain (e.g., Simpson) improve the SRGAN's initial performance on another (e.g., CelebA)? How effectively can fine-tuning adapt the model to a new domain? Can stylized datasets like Simpson help counteract overfitting on real-world datasets that may have inherent biases? With ample data, I can meticulously investigate the influence of various hyperparameters (e.g., learning rate, discriminator complexity, loss function choices) on the model's ability to extract meaningful features across different domains.

The use of image-aligned and pre-partitioned datasets offers substantial benefits. It reduces dataset preparation overhead frees up valuable time for focusing on SRGAN architectures, training strategies, and evaluation. Lastly, predefined splits ensure comparability of results with existing research using these standard datasets.

In the next chapter, these datasets are used with pre-trained weights of VGG19, VGGFace2, EfficientNetB0 for feature extraction while experimenting with hyper-parameters. The use of these datasets with pre-trained weights have reduced execution time significantly. This is a practical and computationally efficient strategy. Bypassing the need to train large feature extractors from scratch drastically reduces compute requirements, enabling more rapid iteration and fine-tuning on specific facial datasets. Pre-training on massive datasets like ImageNet allows these models to develop a rich initial feature representation. Leveraging this knowledge has helped accelerate training and potentially improve the quality of extracted features.

CHAPTER V:

RESULTS AND DISCUSSIONS

5.1 Introduction

This chapter discusses implementation of proposed generative adversarial network models, evaluation metrics and explains results obtained from experiments conducted on three image-aligned datasets with pre-trained network weights. This chapter establishes relationship between research objectives and results. The first part of the chapter discusses various evaluation metrics used for performance evaluation of proposed models. The second part of the chapter focuses on experiments and explanation of results by interpretation of visualizations. The second part of the chapter starts with brief overview of proposed super resolution generative adversarial network models for feature extraction, followed by model hyper-parameters, pre-trained network weights and ends with thorough interpretation of visualization obtained from experiments. Finally, the outcome of various experiments is thoroughly evaluated and analysed to come up with list of recommendations for super resolution generative adversarial networks in feature extraction.

5.2 Evaluation Metrics

An important part of measuring and evaluating super resolution generative adversarial network feature extraction effectiveness is to choose evaluation metrics which are globally accepted and industry recognized. While there are multiple options available to measure feature extraction effectiveness, not all are widely accepted similar to PSNR and SSIM. These are very popular metrics (Horé and Ziou, 2010) used in most super-resolution GAN models for image generation, image identification and feature extraction.

- PSNR is defined as ratio between maximum possible power of signal and power of corrupting noise. In this thesis, models feature extraction effectiveness is evaluated using

PSNR on three image aligned datasets with multiple pre-trained network weights. A high value of PSNR means more noise removed. It is little biased towards over smooth (i.e. blurry) results.

- SSIM is defined as measurement of perceptual difference between two similar images. It considers similarity of edges i.e. high frequency content between denoised and ideal image. It is a better metric compared to PSNR but more complicated to compute. The value range from starts from -1. Higher SSIM numbers represents better structurally aligned images. In this thesis, models feature extraction effectiveness is evaluated using SSIM on three image aligned datasets with multiple pre-trained network weights. SSIM value close to one means image is structurally similar. It is little biased towards over smooth (i.e. blurry) results.

In addition to PSNR and SSIM, generator and discriminator loss are calculated based on loss functions referred in chapter 3, section 3.3.1.

5.3 Interpretation of visualizations

A crucial part in explanation of result is our ability to understand results with ease and gain useful insight. While data and result can be presented in tables, presenting them in visually appealing colourful graph is always better. In this thesis, feature extraction effectiveness for super resolution generative adversarial network is evaluated on PSNR and SSIM metrics. Generator and discriminator loss is evaluated on accuracy for loss functions referred in chapter 3, section 3.3.1. The results of experiments are presented in graphs.

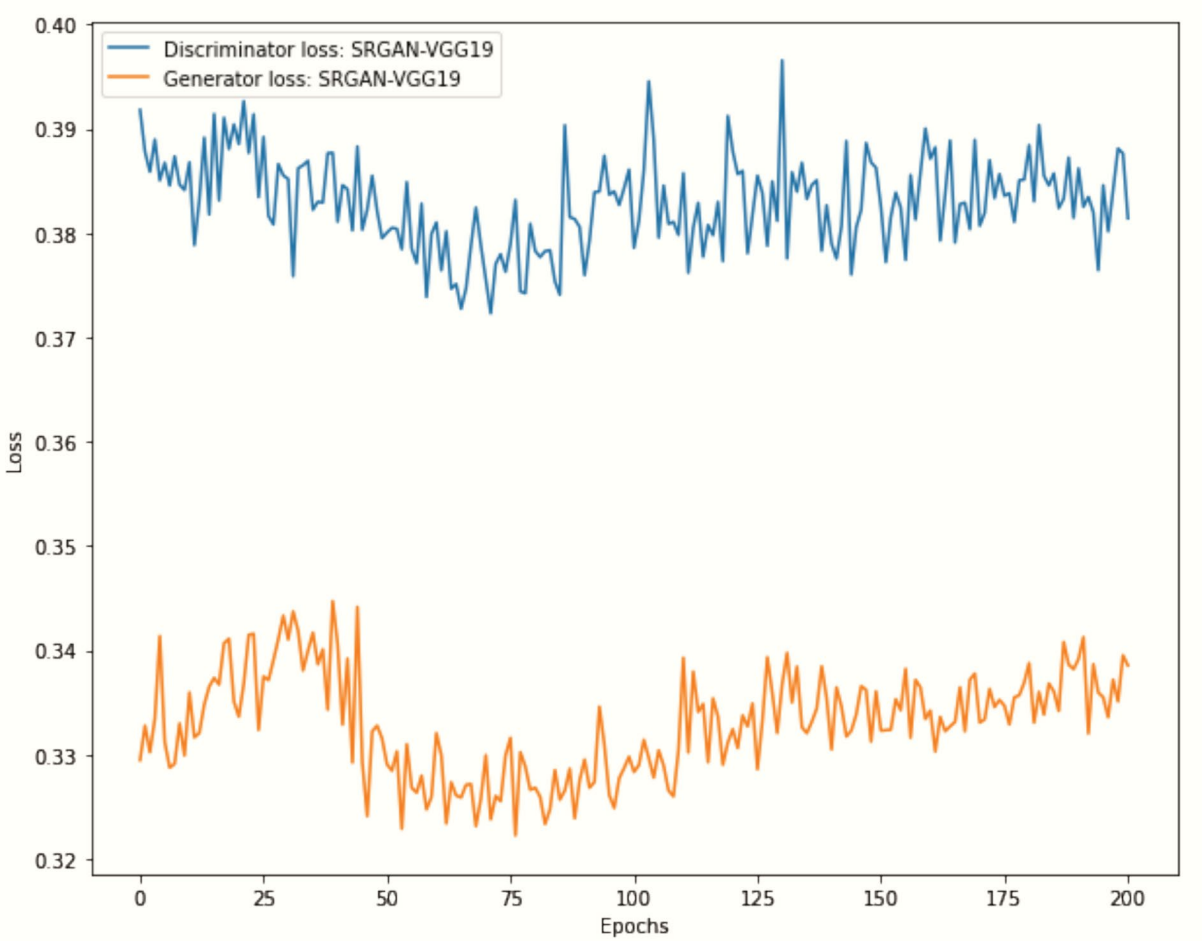


Figure5.1
A sample visualization used in thesis.

The visualization presented in figure 5.1 is an example of visualization used throughout in this chapter. The x-axis represents count of Epochs or iterations used in training loop. The y-axis represents value of graph feature. The visualization presented in figure 5.1, the feature represented is 'Loss'. Each line on x-y axis represents experiment outcome and each line is coloured differently to easily distinguish it from other lines. Graph corners are used to tag name for each line. This helps in overall understanding of visualization.

5.4 Evaluation of feature extraction models and results

This thesis uses SRGAN (Ledig et al., 2017) as a base GAN model to perform experiments on three image aligned datasets for feature extraction. The hypothesis is about substituting feature extractor by VGG19, VGGFace2 and EfficientNet-B0 in SRGAN. The conceptual architecture of SRGAN model used in thesis is listed in figure 5.2.

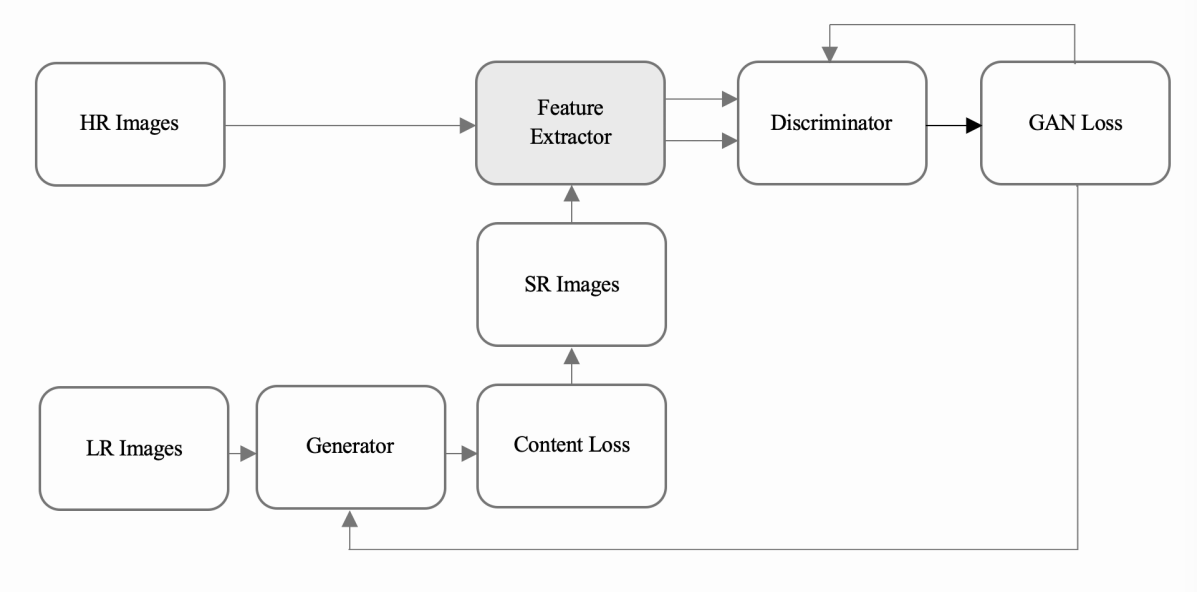


Figure5.2
Conceptual SRGAN model architecture.

To meet stated thesis objectives, more than nine experiments have been performed. These experiments use three image aligned datasets with four pre-trained network weights to evaluate feature extraction performance and to evaluate loss for Generator and Discriminator. The evaluation metrics used to measure performance of feature extraction are PSNR and SSIM. The evaluation metric used to measure loss function outcome for Generator and Discriminator is MSE loss with binary_crossentropy.

5.4.1 Hyper-parameters

The importance of hyper-parameters in an experiment is paramount. By carefully testing and choosing right value for hyper-parameters, this thesis managed to reduce experiment run time significantly. Despite using less than 500 epochs for each experiment, this thesis managed to equal and, in some cases, outperform default VGG based SRGAN feature extraction results.

Table 5.1 neatly consolidate and summarize common hyper-parameters used in each experiment and lists down changes compared to original SRGAN implementation.

Experiments defaults	SRGAN original defaults
Epochs = 51, 101, 151, 201 and 401 Batch size = 8 Optimizer = Adam with learning rate 0.0002 and β_1=0.5 Low-resolution image shape = 64x64x3 High-resolution image shape = 256x256x3 Momentum = 0.8 Generator activation function = ReLU Discriminator activation function = LeakyReLU with α = 0.2 Generator output activation = Tanh Discriminator output activation = Sigmoid	Epochs = 10000 and 100000 Network weight = VGG Optimizer = Adam with learning rate 0.0001 and β_1=0.9 and learning rate 0.00001 and β_1=0.9 Generator activation function = PReLU Discriminator activation function = LeakyReLU with α = 0.2 Discriminator output activation = Sigmoid

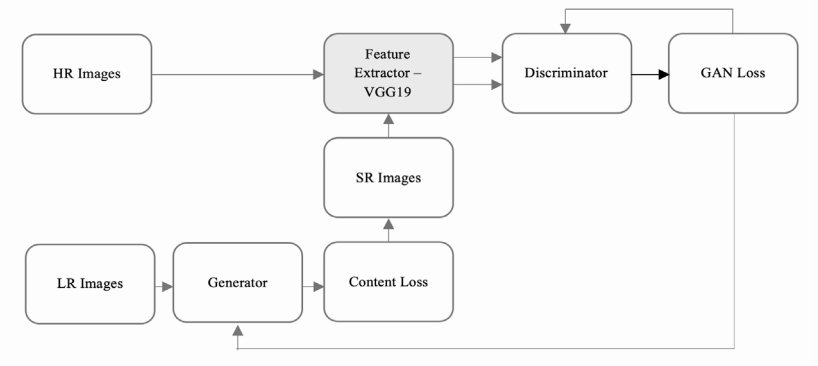
Table 5.1 List of common hyper-parameters

The thesis uses common hyper-parameters listed in table 5.1 in every experiments. Hyper-parameters specific to an experiment are listed in separate table at respective section. For example, hyper-parameters specific to experiments conducted on CelebA dataset are listed in table 5.2.

5.4.2 Model evaluation on CelebA dataset and results

Image aligned version of CelebA dataset is a very large-scale complex face images dataset with numerous variations. Thus, it is an ideal candidate dataset for feature extraction work. This thesis conducted three experiments on CelebA dataset. These experiments are focused on feature extraction module by substituting default feature extractor with VGG19, VGGFace2 and EfficientNetB0 in SRGAN. Additionally, several experimental changes are made to default SRGAN hyper-parameters. These changes are listed in table 5.1.

Table 5.2 neatly consolidates and summarizes three experiments along with respective model architectures, hyper-parameters and pre-trained network weights for CelebA dataset.

Dataset name	CelebA dataset
Experiment 1: SRGAN-VGG19	Model architecture = Use VGG19 feature extractor  Pre-trained network weight = VGG19 Imagenet VGG19 Imagenet loss function = MSE for Accuracy metrics Adversarial loss = binary_crossentropy

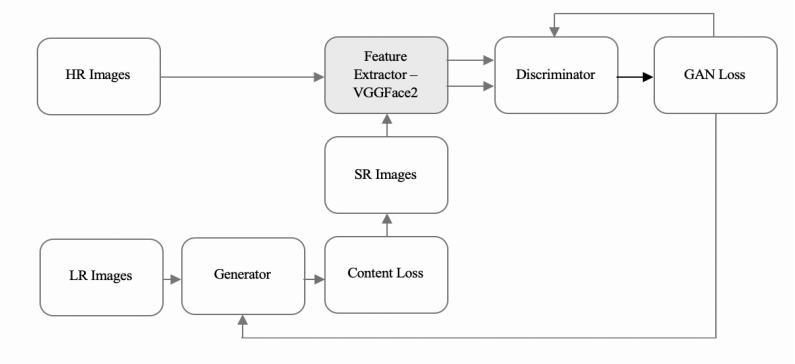
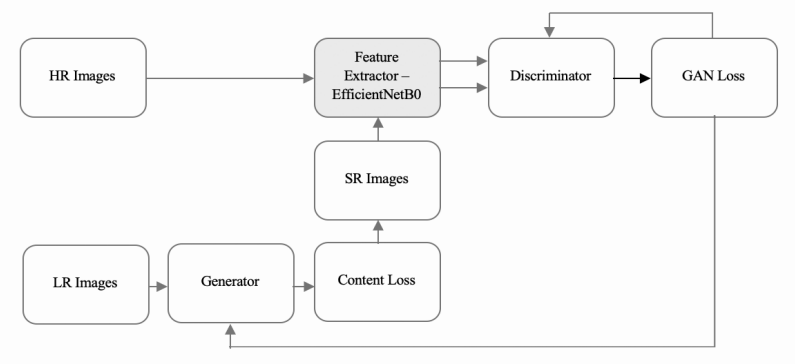
	Epochs = 51, 101, 151 and 201
Experiment 2: SRGAN-VGGFace2	Model architecture = Use VGGFace2-VGG16 feature extractor  Pre-trained network weight = VGGFace2-VGG16 VGGFace2-VGG16 loss function = MSE for Accuracy metrics Adversarial loss = binary_crossentropy Epochs = 51, 101, 151 and 201
Experiment 3: SRGAN-EfficientNetB0	Model architecture = Use EfficientNetB0 feature extractor  Pre-trained network weight = EfficientNetB0 Imagenet EfficientNetB0 Imagenet loss function = MSE for Accuracy metrics Adversarial loss = binary_crossentropy Epochs = 51, 101, 151 and 201

Table 5.2 CelebA dataset experiments details

Experiments listed in table 5.2 meet the objective and scope of this thesis stated in chapter 1, section 1.3 and chapter 1, section 1.4. Let's quickly recall and summarize scope of the thesis.

- Successful use of VGG19, VGGFace2 and EfficientNet as pre-trained transfer learning backbones.
- Development of new GAN based super-resolution architecture by changing feature extractor and changing various hyper parameters.
- Evaluate model performance using Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM) metrics on various large challenging face image datasets.

In the next section, we aim to complete thorough interpretation of visualization obtained through experiments listed in table 5.2.

5.4.2.1 Generator and Discriminator loss analysis

One of the fundamental requirements in evaluation of generative adversarial network performance is to evaluate Generator and Discriminator loss values. We conducted experiments on CelebA dataset which are focused on feature extraction module by substituting default feature extractor with VGG19, VGGFace2 and EfficientNetB0, along with tuning of default hyper-parameters in SRGAN. The figure 5.3 represents Generator and Discriminator loss values obtained after experiments.

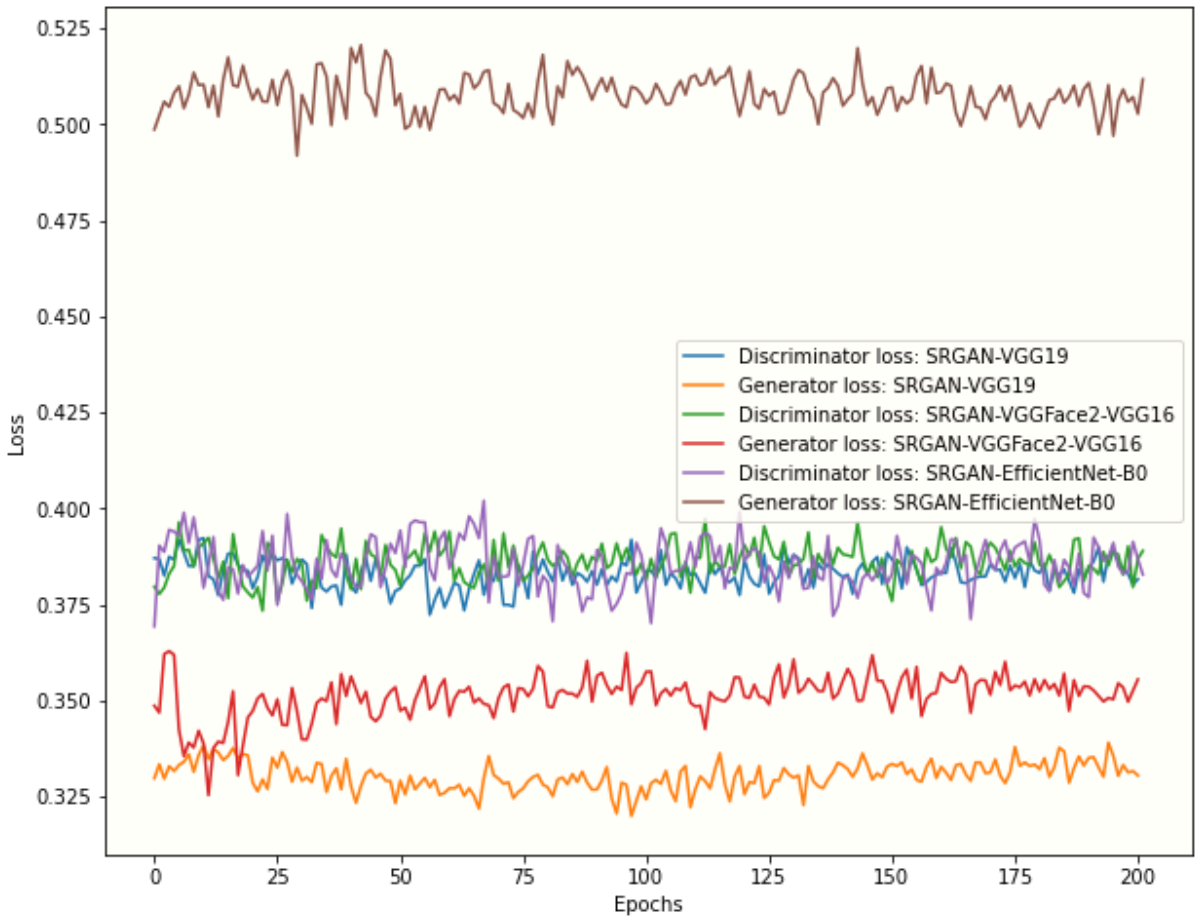


Figure5.3
Generator and Discriminator loss value analysis for CelebA dataset.

It is evident from figure 5.3 that our experimental changes in hyper-parameters and use of pre-trained network weights results in very stable and low (Brownlee, 2021) Generator loss for VGG19 pre-trained weights. The Discriminator loss is low and stability is rock solid for all pre-trained weights. These results are much lower and stable compared to standard SRGAN implementation.

5.4.2.2 PSNR and SSIM analysis

Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index (SSIM) metrics are industry benchmark metrics for evaluating generative adversarial network performance. This is also one of the thesis objectives. SSIM is a much better evaluation metrics compared to PSNR for generative adversarial networks. We have evaluated model performance on both metrics to maintain parity with industry practices. PSNR and SSIM values are plotted for experiments listed in table 5.2.

- Full images analysis: It is evident from figure 5.4a and 5.4b that VGGFace2-VGG16 has stable PSNR values while VGG19 has rapidly increasing PSNR values. At higher epochs, VGG19 should outperform other pre-trained weights for PSNR values. On the contrary, VGGFace2-VGG16 performs better on SSIM metrics and shows an increasing trend at higher epochs.

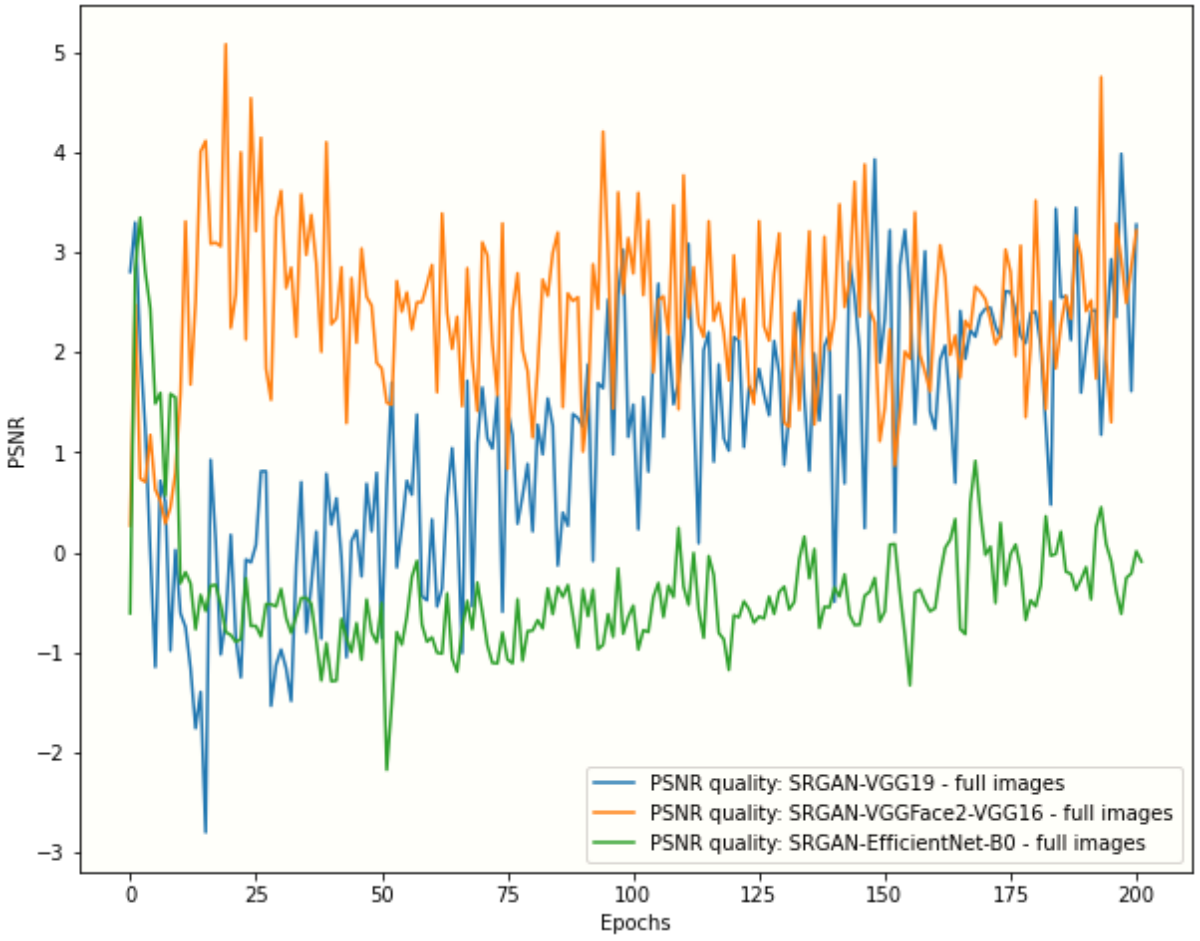


Figure5.4a
PSNR value analysis on full image for CelebA dataset.

Figure5.4b
SSIM value analysis on full image for CelebA dataset.

- Feature extraction analysis – face extraction: It is visible from figure 5.5a and 5.5b that VGGFace2-VGG16 has stable PSNR values while VGG19 and EfficientNetB0 have increasing PSNR values. At higher epochs, VGG19 should outperform other pre-trained weights for PSNR values. The SSIM metric value echo similar outcome.

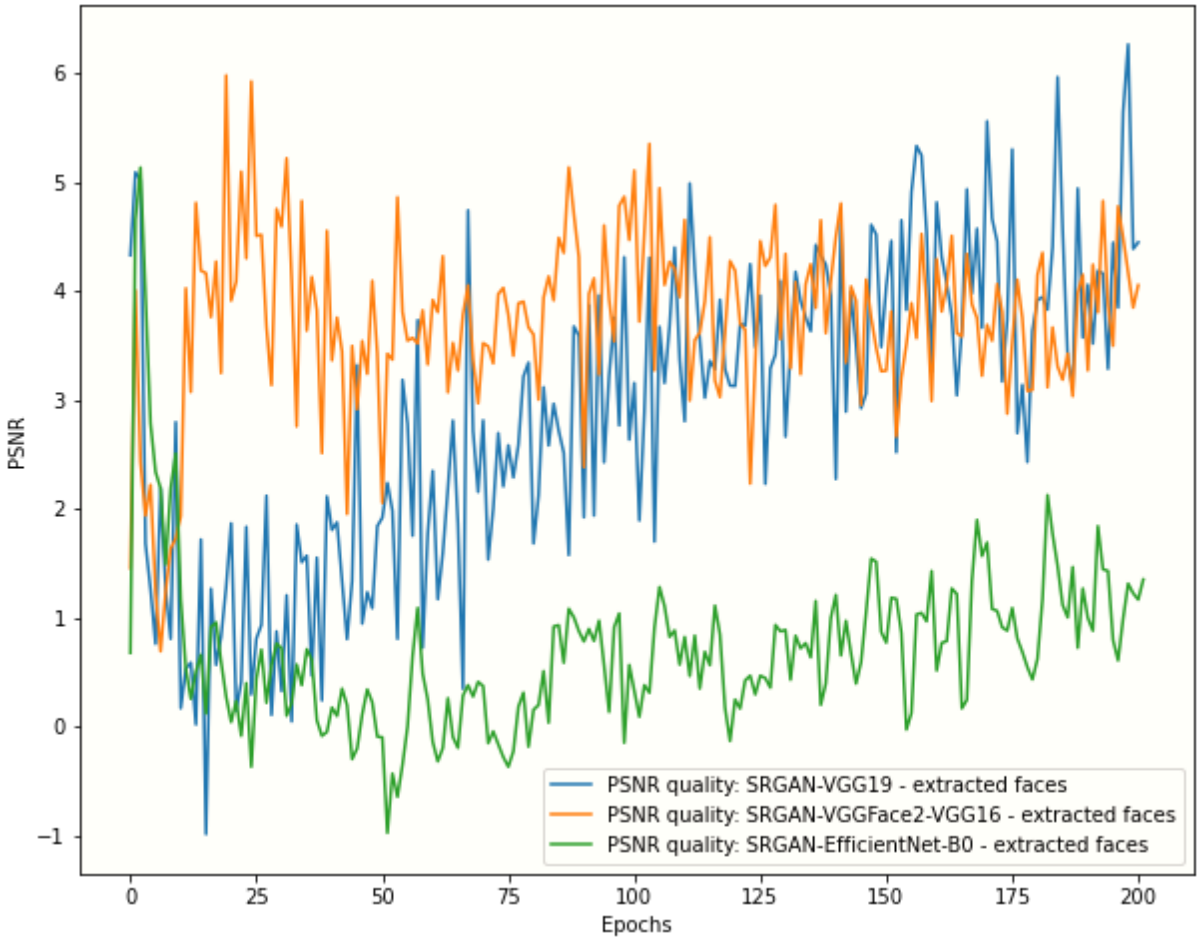


Figure5.5a
PSNR value analysis on extracted faces for CelebA dataset.

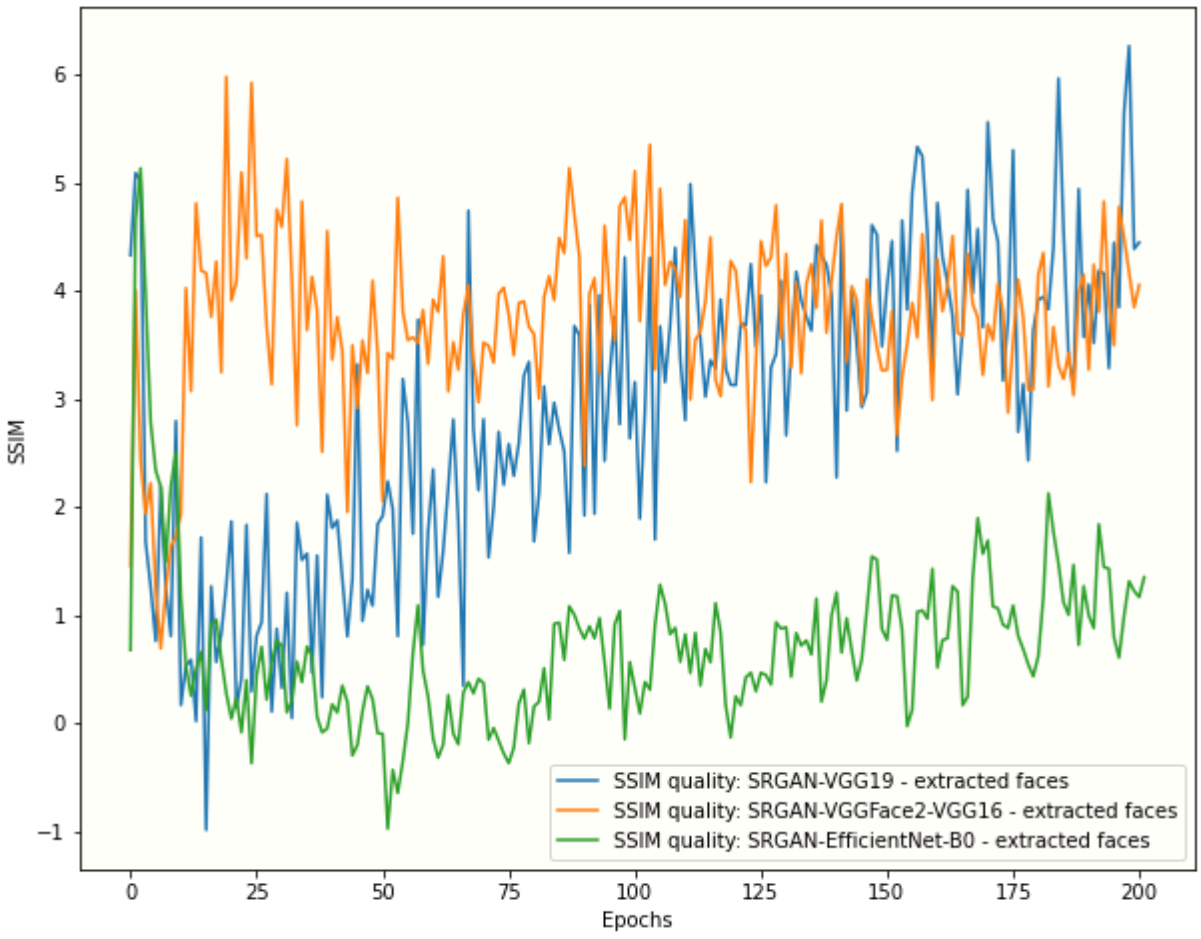


Figure5.5b
SSIM value analysis on extracted faces for CelebA dataset.

- Feature extraction analysis – right eye extraction: It is clear from figure 5.6a and 5.6b that VGGFace2-VGG16 has stable PSNR values while VGG19 and EfficientNetB0 have increasing PSNR values. At higher epochs, VGG19 and EfficientNetB0 should outperform VGGFace2-VGG16 for PSNR values. The SSIM metrics echo similar outcome.

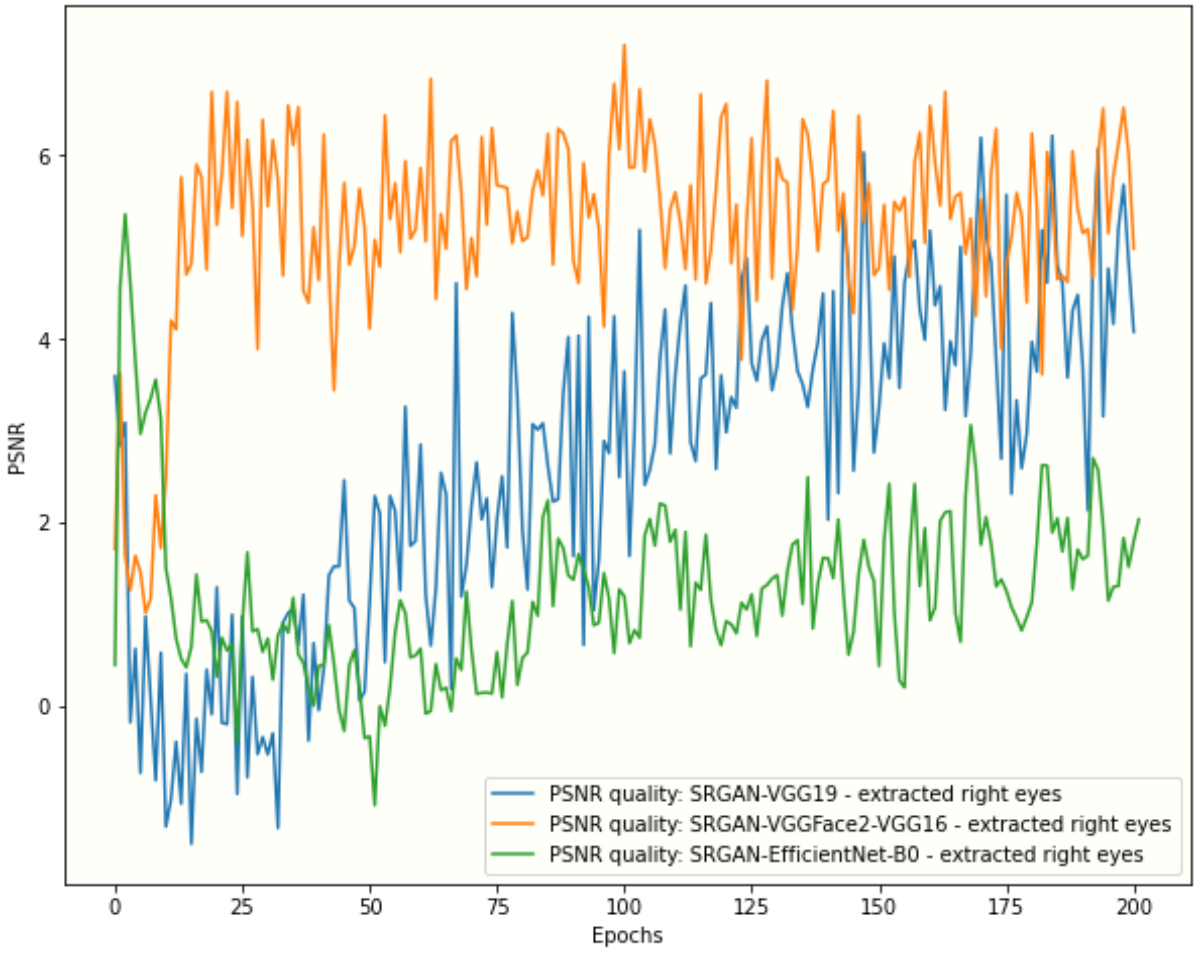


Figure5.6a
PSNR value analysis on extracted right eye for CelebA dataset.

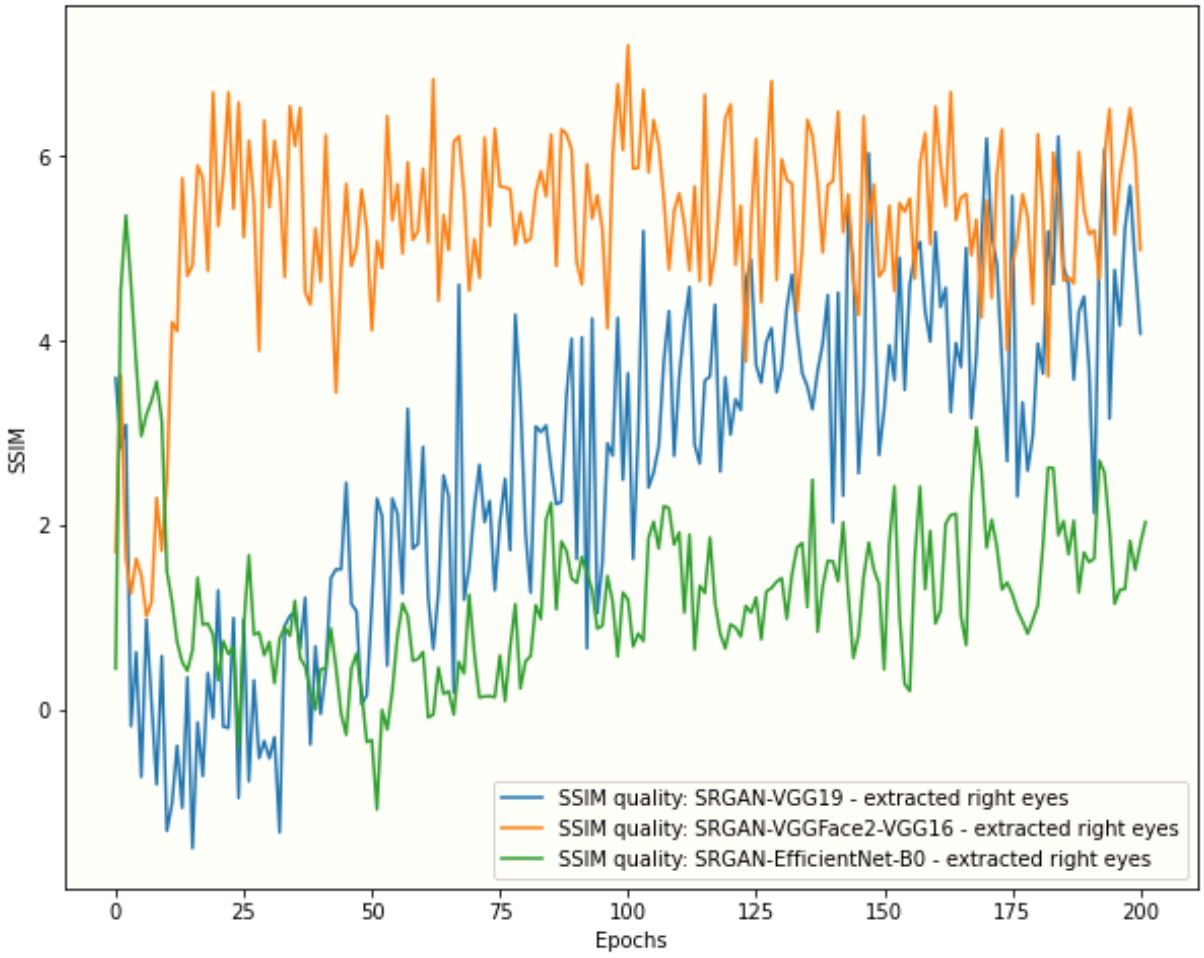


Figure5.6b
SSIM value analysis on extracted right eye for CelebA dataset.

- **Feature extraction analysis – left eye extraction:** It is quite evident from figure 5.7a and 5.7b that VGGFace2-VGG16 has stable PSNR values while VGG19 and EfficientNetB0 have increasing PSNR values. At higher epochs, VGG19 and EfficientNetB0 should outperform VGGFace2-VGG16 for PSNR values. The SSIM metrics echo similar outcome.

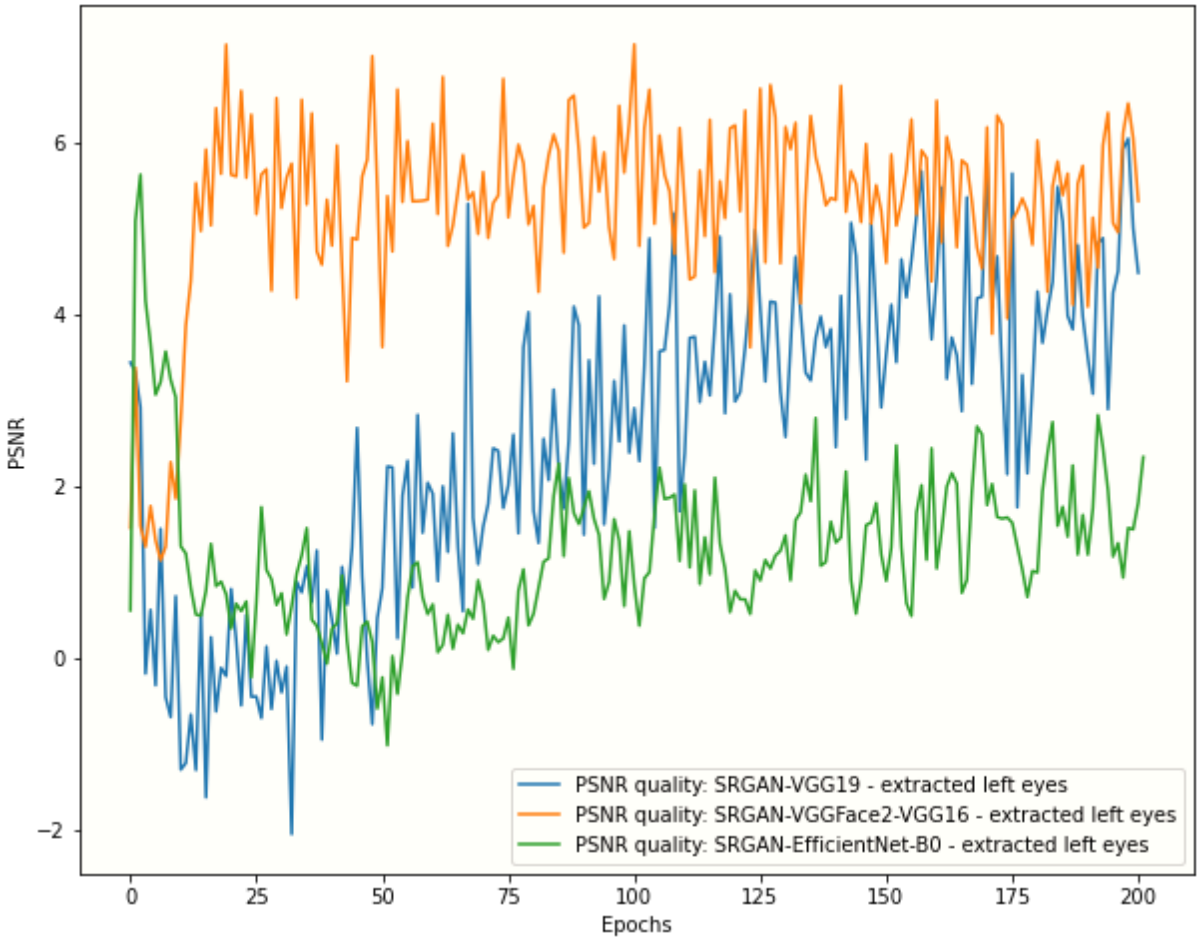


Figure5.7a
PSNR value analysis on extracted left eye for CelebA dataset.

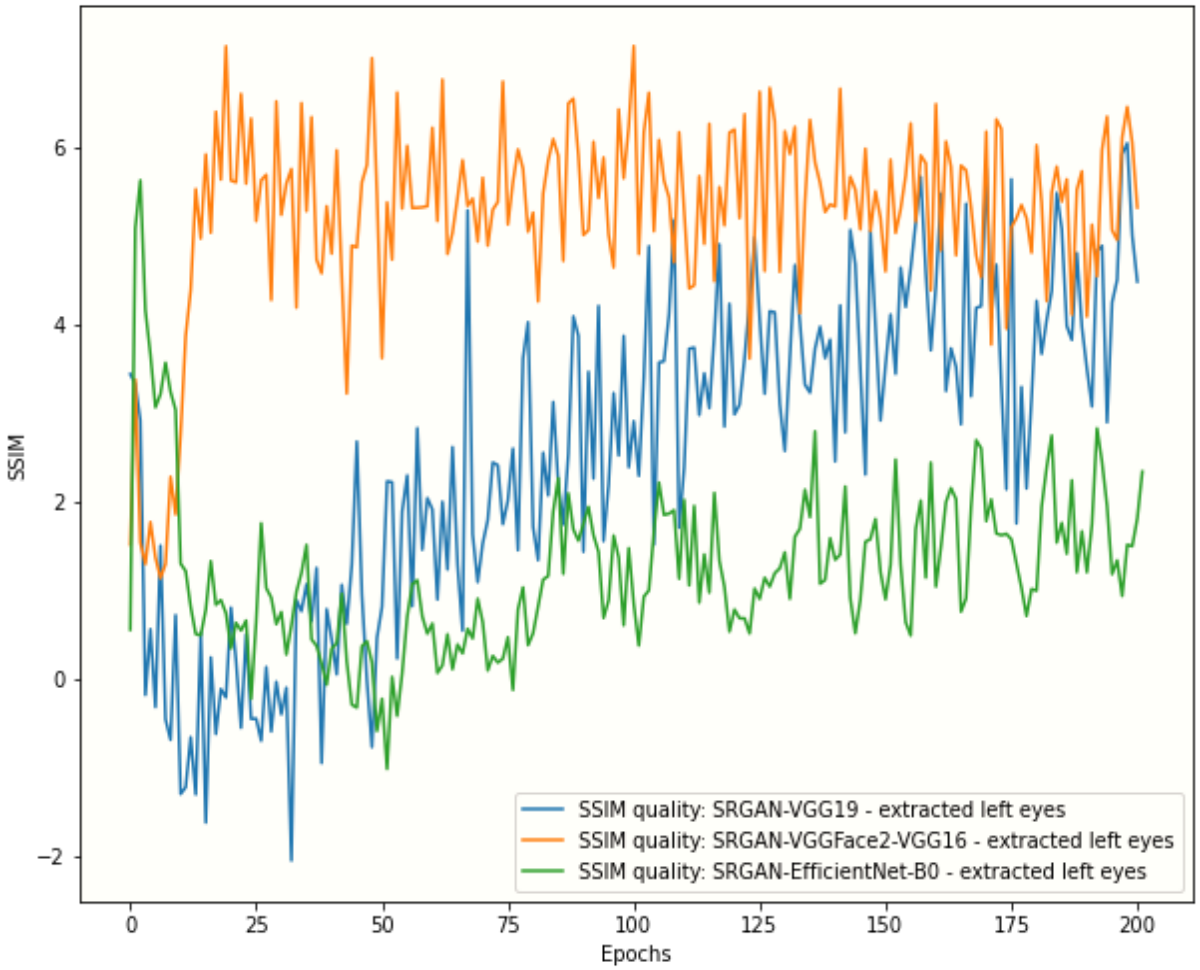


Figure5.7b
SSIM value analysis on extracted left eye for CelebA dataset.

5.4.2.3 CelebA conclusion

Experiments conducted on CelebA dataset showcase clearly that selection of custom hyper-parameter values with pre-trained learning network backbone have resulted in remarkably low level of loss (Brownlee, 2021) for Generator and Discriminator. The loss is stable and predictable. This is a great achievement considering epochs count is in range of sub-300. This is remarkable compared to other implementations based on SRGAN (Ledig et al., 2017; Li et al., 2020; Zhang et al., 2020) where epochs count is 10^5 or even more.

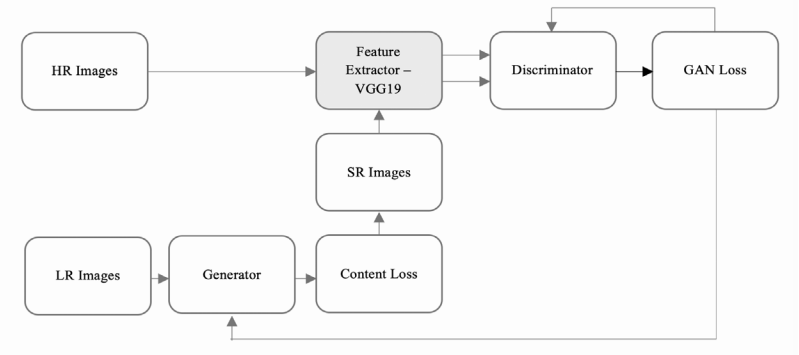
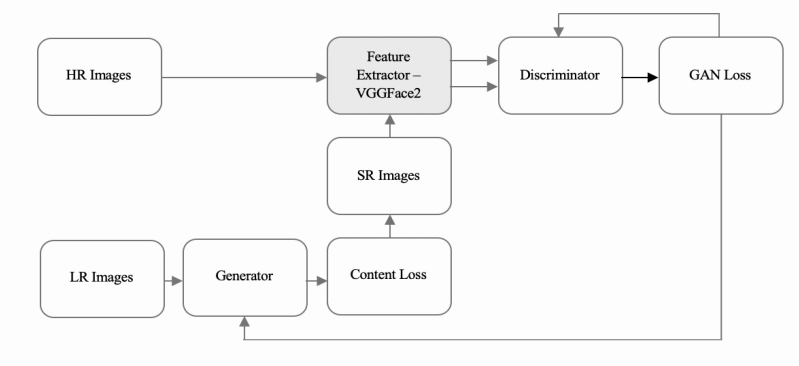
Experiments conducted for feature extractions shows great results as well. The use of pre-trained learning network backbone shows that by using an already learned network with appropriate hyper-parameter values, an increasingly high level of effectiveness can be maintained in feature extraction, even with sub 300 number of epochs. Out of three pre-trained weights, VGG19 and EfficientNetB0 shows great results for feature extraction. The PSNR and SSIM number could have been very high if epochs counts could be 10^5 or more, together with high batch size which couldn't be achieved given memory constraints on lab hardware.

5.4.3 Model evaluation on LFW dataset and results

Image aligned deep funnelled version of LFW dataset is a large-scale complex face images dataset with numerous variations. Thus, it is an ideal candidate dataset for feature extraction work. This thesis conducted three experiments on LFW dataset. These experiments are focused on feature extraction module by substituting default feature extractor with VGG19, VGGFace2 and EfficientNetB0 in SRGAN. Additionally, several experimental changes are made to default SRGAN hyper-parameters. These changes are listed in table 5.1.

Table 5.3 neatly consolidates and summarizes three experiments along with respective model architectures, hyper-parameters and pre-trained network weights for LFW dataset.

Dataset name	LFW dataset
Experiment 1: SRGAN-VGG19	Model architecture = Use VGG19 feature extractor

	 Pre-trained network weight = VGG19 Imagenet VGG19 Imagenet loss function = MSE for Accuracy metrics Adversarial loss = binary_crossentropy Epochs = 51, 101, 151 and 201
Experiment 2: SRGAN-VGGFace2	Model architecture = Use VGGFace2-VGG16 feature extractor  Pre-trained network weight = VGGFace2-VGG16 VGGFace2-VGG16 loss function = MSE for Accuracy metrics Adversarial loss = binary_crossentropy Epochs = 51, 101, 151 and 201
Experiment 3: SRGAN-EfficientNetB0	Model architecture = Use EfficientNetB0 feature extractor

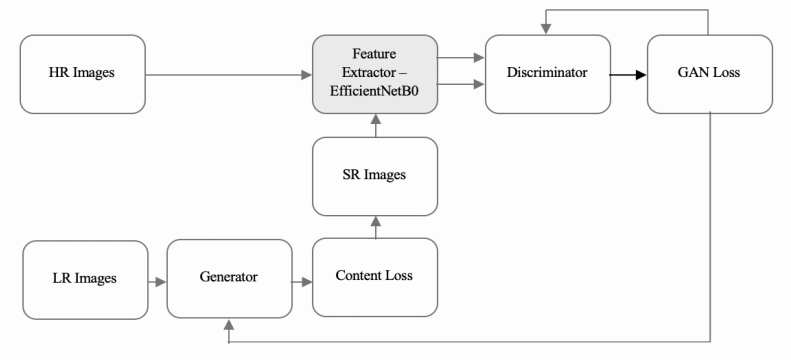
	 Pre-trained network weight = EfficientNetB0 Imagenet EfficientNetB0 Imagenet loss function = MSE for Accuracy metrics Adversarial loss = binary_crossentropy Epochs = 51, 101, 151 and 201

Table 5.3 LFW dataset experiments details

Experiments listed in table 5.3 meet the objective and scope of this thesis stated in chapter 1, section 1.3 and chapter 1, section 1.4. In the next section, we run through results by performing thorough interpretation of visualization obtained through experiments.

5.4.3.1 Generator and Discriminator loss analysis

One of the fundamental requirements in evaluation of generative adversarial network performance is to evaluate Generator and Discriminator loss values. The experiments conducted on LFW dataset are focused on feature extraction module by substituting default feature extractor with VGG19, VGGFace2 and EfficientNetB0, along with tuning of default hyper-parameters in SRGAN. The figure 5.8 represents Generator and Discriminator loss values obtained after experiments.

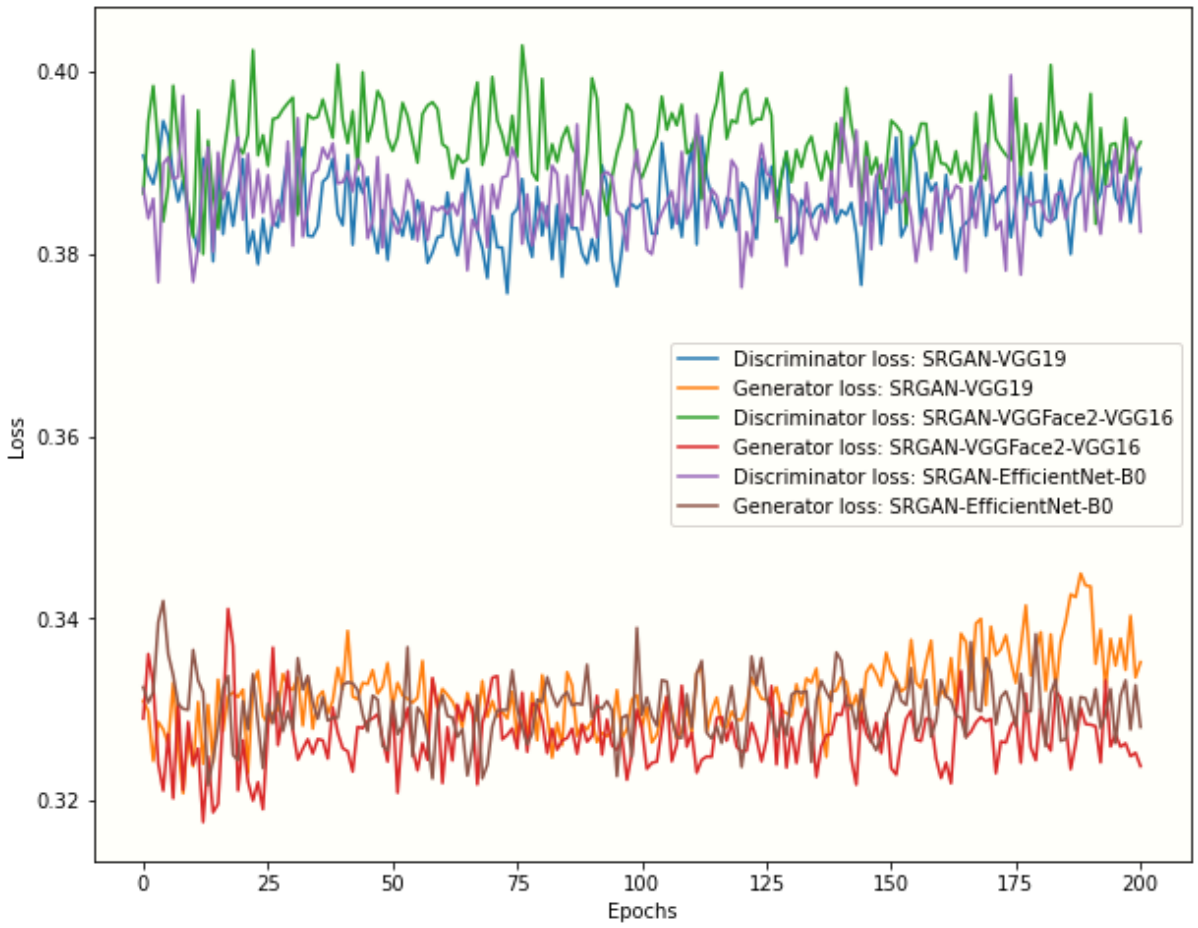


Figure5.8
Generator and Discriminator loss value analysis for LFW dataset.

It is clearly visible from figure 5.8 that we are able to standardize Generator loss across all pre-trained learning backbones compared to CelebA experiments where Generator loss was different for every pre-trained learning backbone. Our experimental changes in hyper-parameters and use of pre-trained network weights results in very stable and low (Brownlee, 2021) Generator loss for all pre-trained learning network weights. The Discriminator loss is low and very stable for all pre-trained learning weights. These results are much lower and stable compared to standard SRGAN implementation.

5.4.3.2 PSNR and SSIM analysis

We have evaluated model performance on both metrics to maintain parity with industry practices. PSNR and SSIM values are plotted for experiments listed in table 5.3.

- **Full images analysis:** It is visible from figure 5.9a and 5.9b that VGGFace2-VGG16 and EfficientNetB0 have stable PSNR values while VGG19 has an overall upward trajectory for PSNR values. At higher epochs, VGG19 should outperform other pre-trained weights for PSNR values. VGGFace2-VGG16 and VGG19 performs better on SSIM metrics and show an upward trajectory in-line with increasing number of epochs.

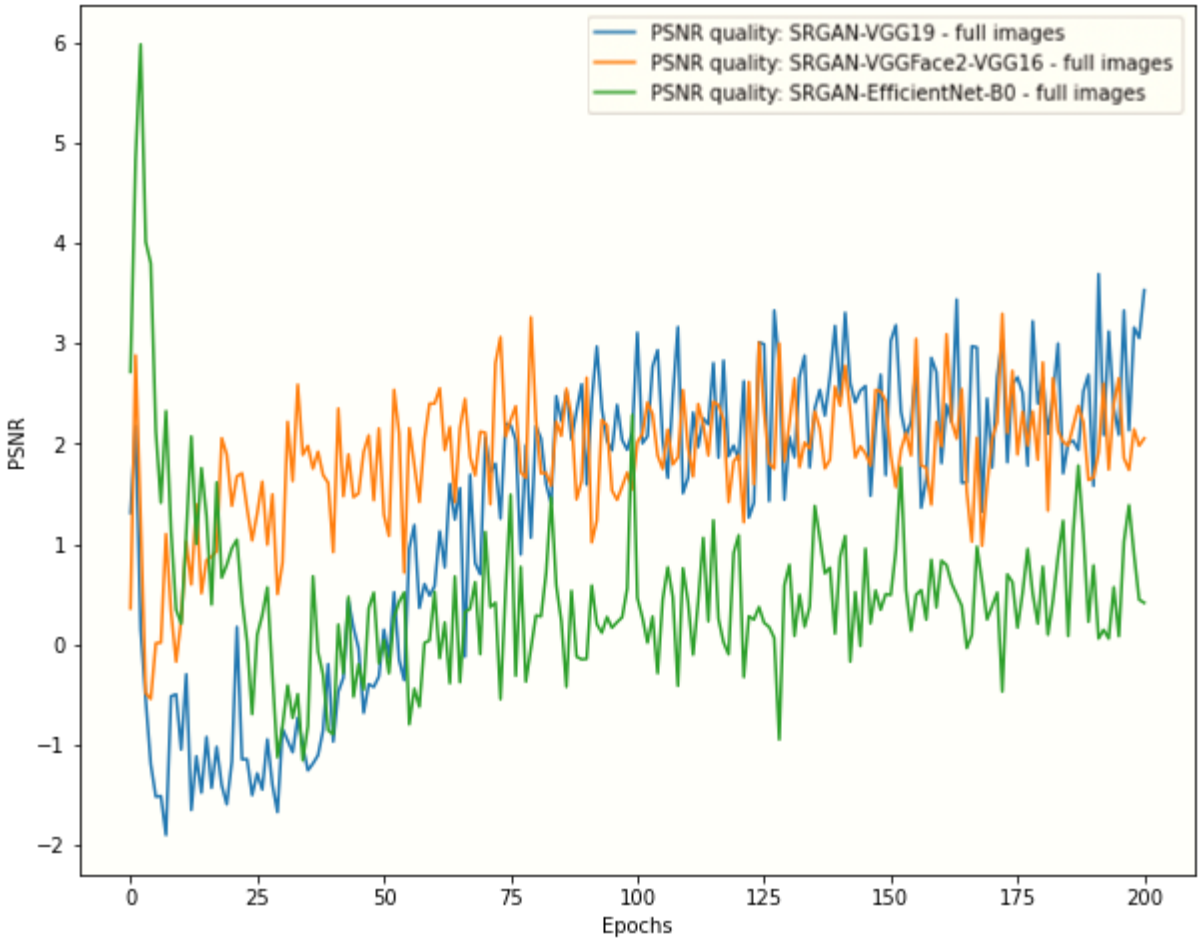


Figure5.9a
PSNR value analysis on full image for LFW dataset.

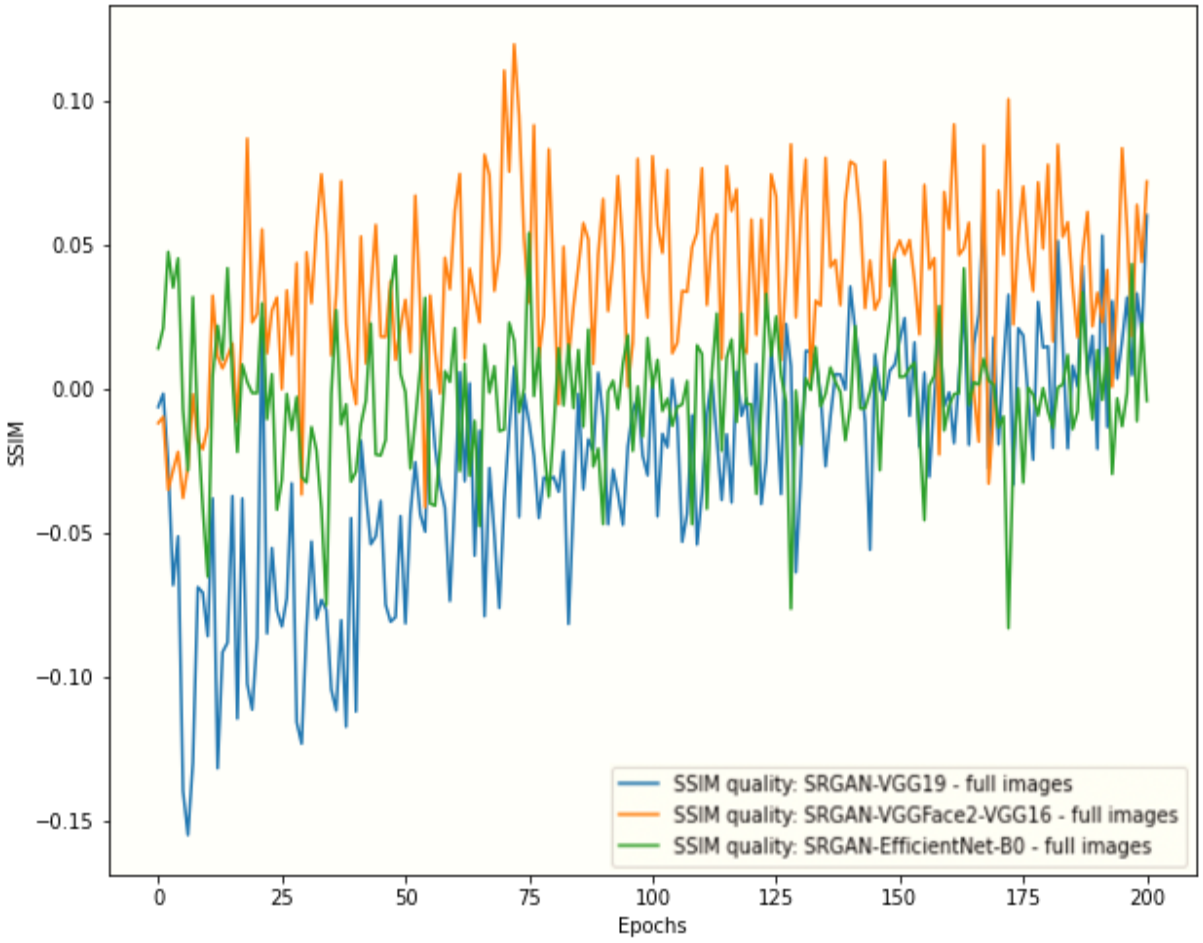


Figure5.9b
SSIM value analysis on full image for LFW dataset.

- **Feature extraction analysis – face extraction:** It is evident from figure 5.10a and 5.10b that VGGFace2-VGG2 and EfficientNetB0 have stable PSNR values while VGG19 shows an upward trajectory of PSNR values, in-line with increasing number of epochs. At higher epochs, VGG19 should outperform other pre-trained weights for PSNR values. The SSIM metric value echo similar outcome.

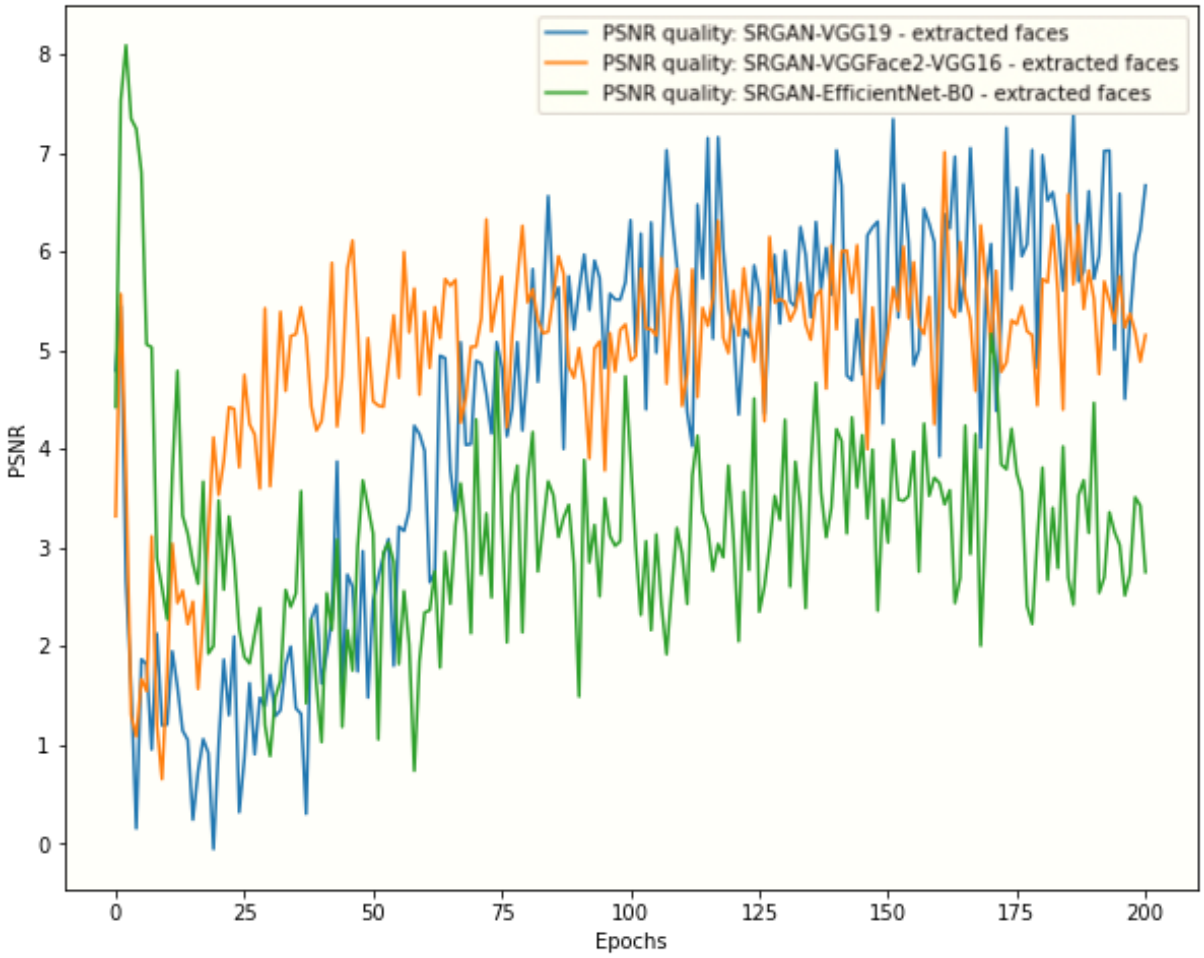


Figure5.10a
PSNR value analysis on extracted faces for LFW dataset.

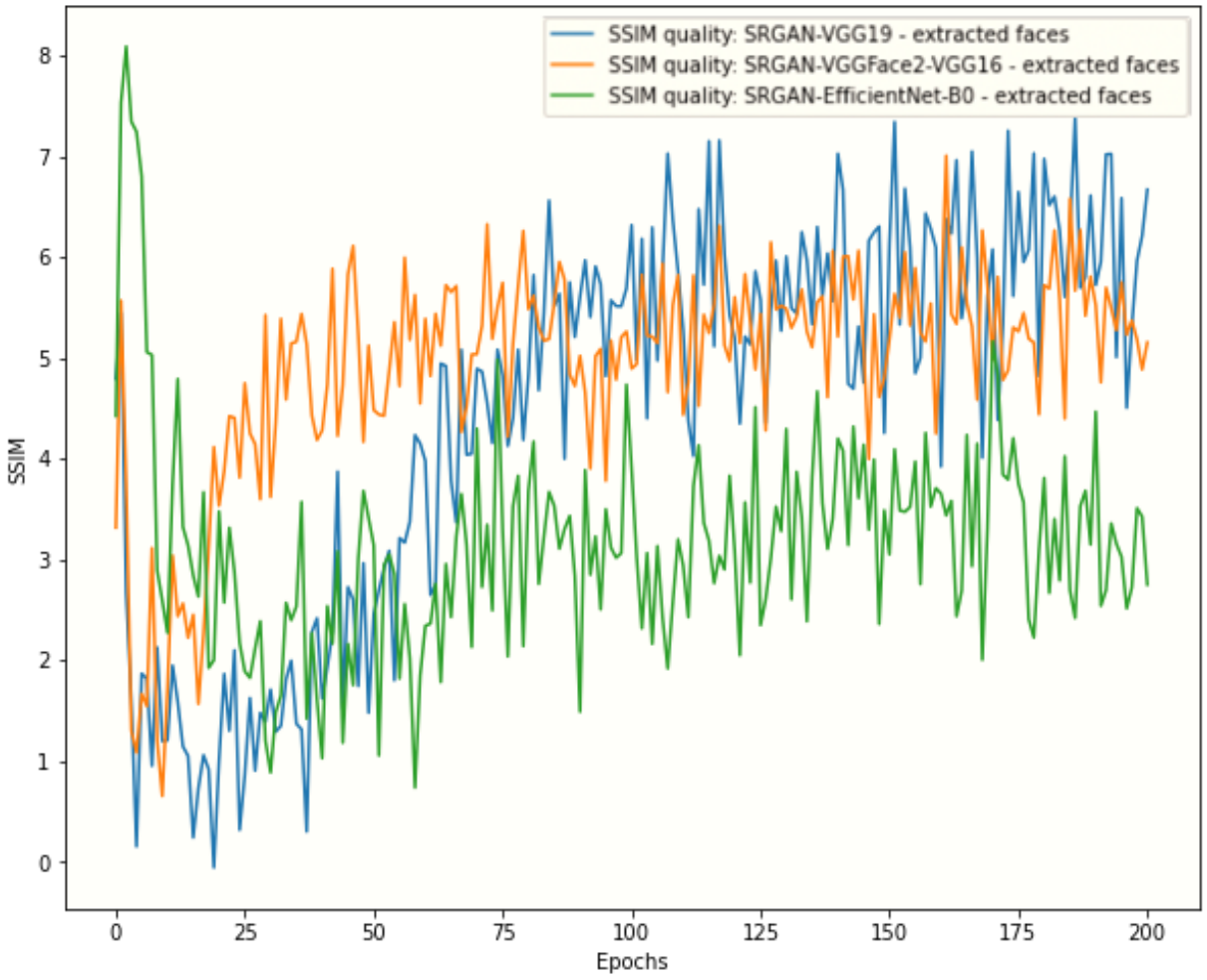


Figure5.10b
SSIM value analysis on extracted faces for LFW dataset.

- **Feature extraction analysis – right eye extraction:** It is visible from figure 5.11a and 5.11b that VGGFace2-VGG16 and EfficientNetB0 have slow upward trajectory of PSNR values while VGG19 fast upward trajectory of PSNR values, in-line with increasing number of epochs. At higher epochs, VGG19 should outperform other pre-trained weights for PSNR values. The SSIM metrics echo similar outcome.

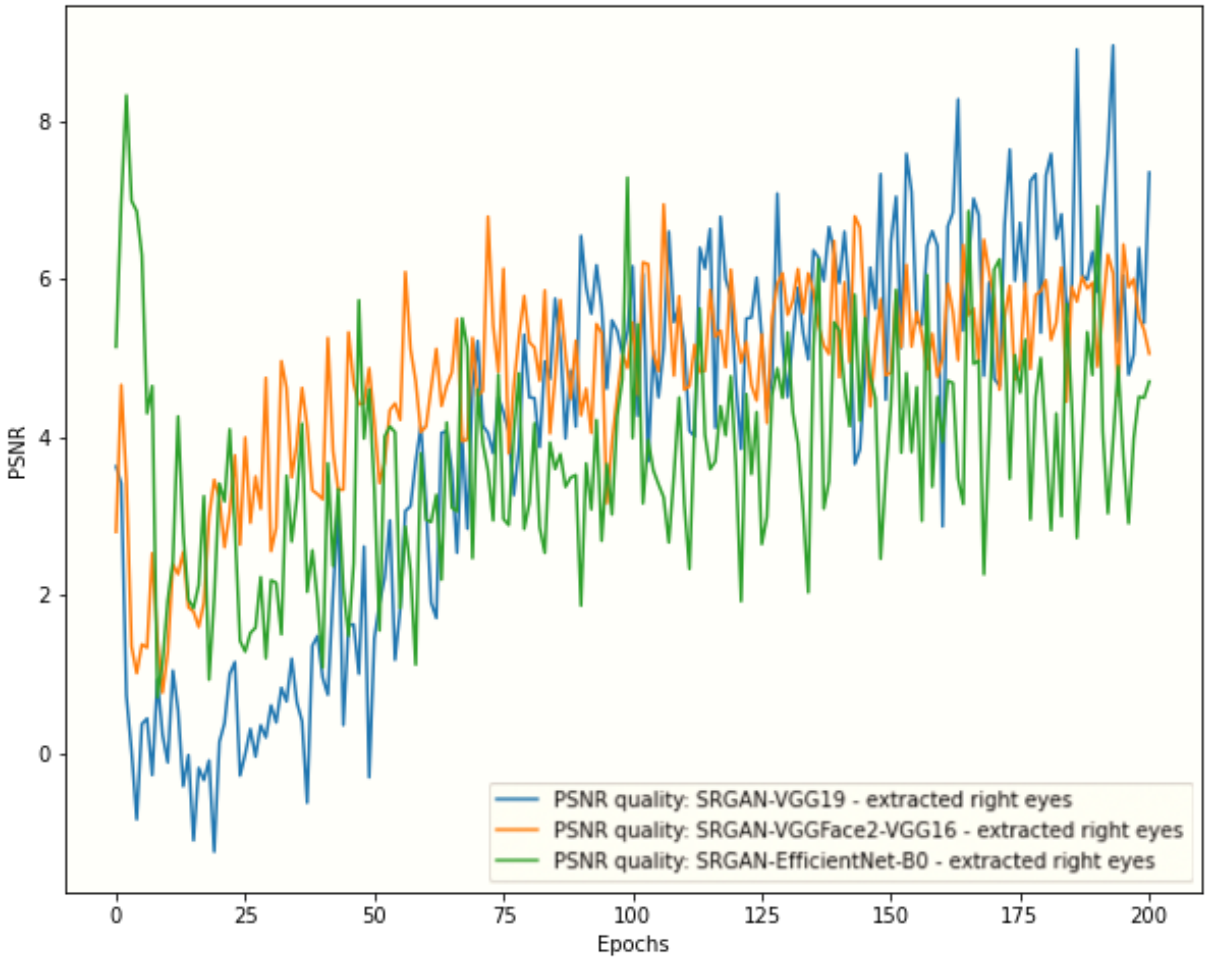


Figure5.11a
PSNR value analysis on extracted right eye for LFW dataset.

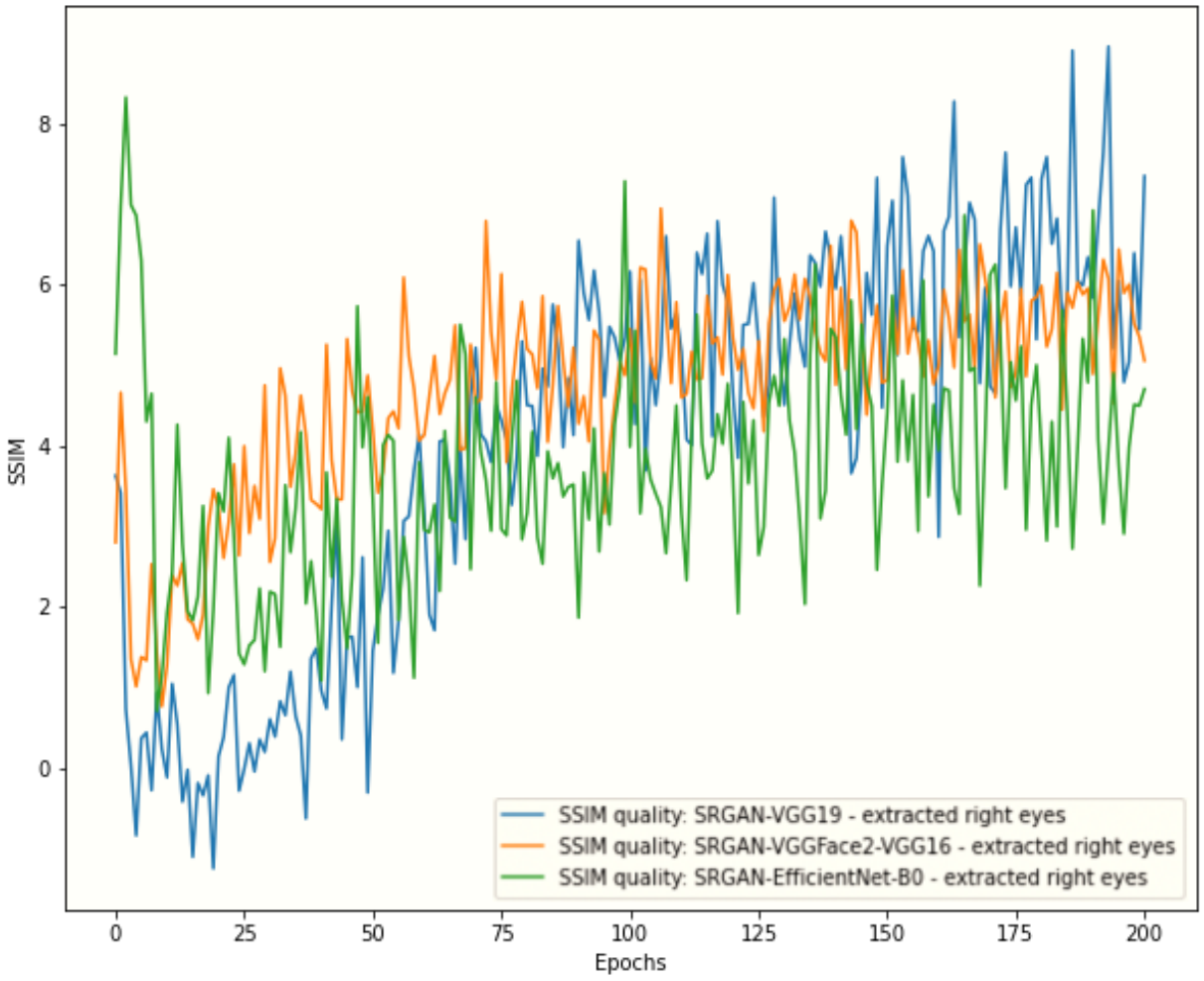


Figure5.11b
SSIM value analysis on extracted right eye for LFW dataset.

- **Feature extraction analysis – left eye extraction:** It is clear from figure 5.12a and 5.12b that all pre-trained leaning networks have upward trajectory of PSNR values, in-line with increasing number of epochs count. However, VGG19 shows steeper growth. The SSIM metrics echo similar outcome, ever so slightly for VGGFace2-VGG16 and EfficientNetB0.

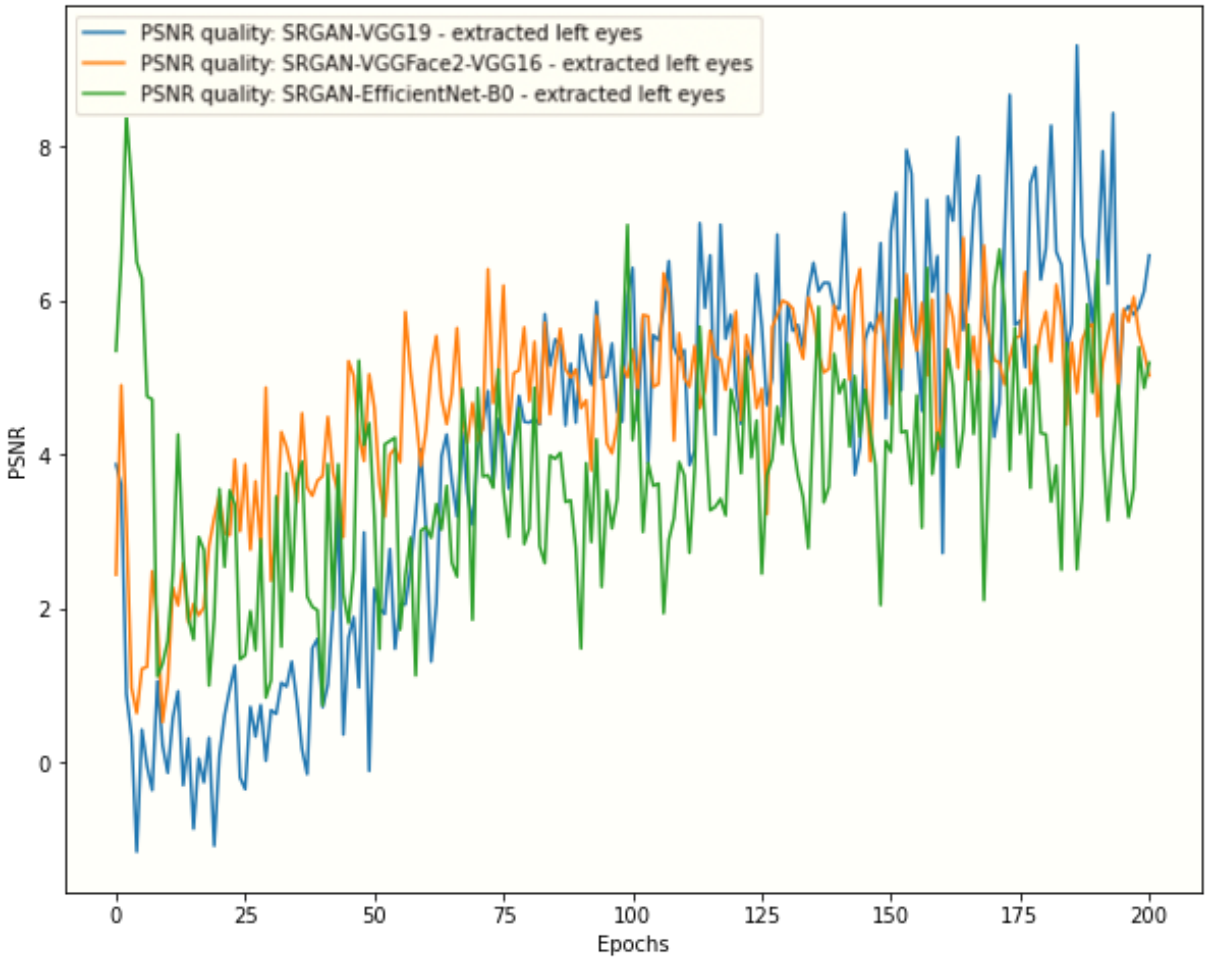


Figure5.12a
PSNR value analysis on extracted left eye for LFW dataset.

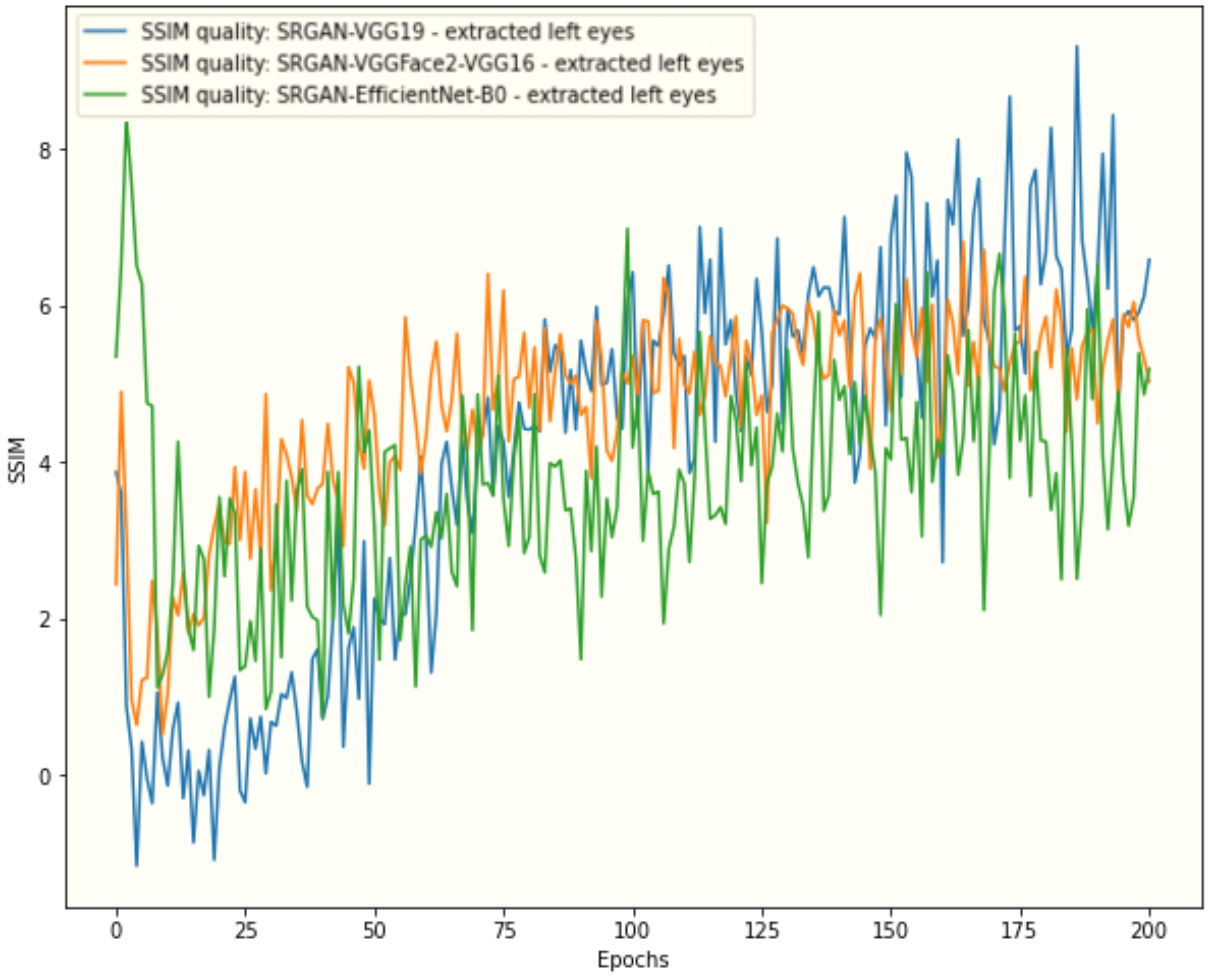


Figure5.12b
SSIM value analysis on extracted left eye for LFW dataset.

5.4.3.3 LFW conclusion

Experiments conducted on LFW dataset clearly shows that we are able to standardize Generator loss across all pre-trained learning backbones compared to CelebA experiments where Generator loss was different for every pre-trained learning backbone. It reaffirms our objectives that selection of custom hyper-parameter values with pre-trained learning network backbone have resulted in remarkably stable and low level of loss (Brownlee, 2021) for Generator and Discriminator. The loss is stable and predictable considering epochs count is in range of sub-300.

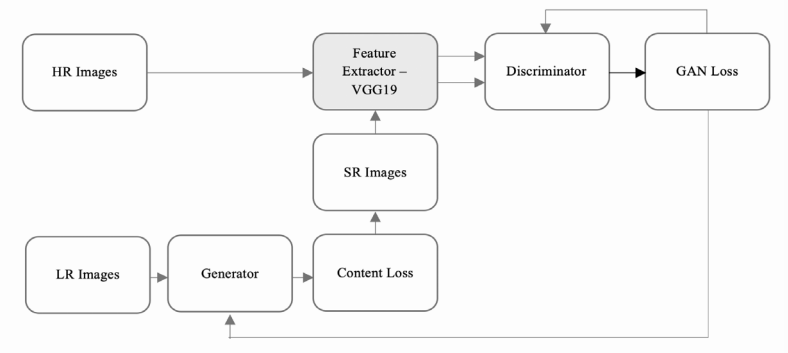
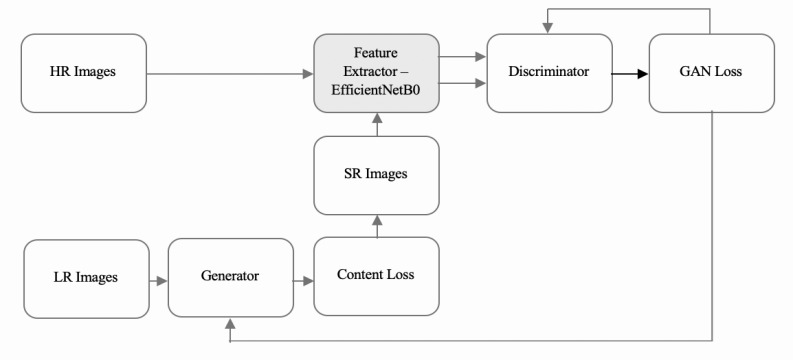
This is remarkable compared to other implementations based on SRGAN (Ledig et al., 2017; Li et al., 2020; Zhang et al., 2020) where epochs count is 10^5 or even more. Experiments conducted for feature extractions shows good results as well. The use of pre-trained learning network backbone shows that by using an already learned network with appropriate hyper-parameter values, an increasingly high level of effectiveness can be maintained in feature extraction, even with sub 300 number of epochs.

Out of three pre-trained weights, VGG19 shows great potential for feature extraction, in-line with increasing number of epochs. The PSNR and SSIM number could have been very high if epochs counts could be 10^5 or more, together with high batch size which couldn't be achieved given memory constraints on lab hardware.

5.4.4 Model evaluation on Simpson dataset and results

Image aligned version of Simpson dataset is a very large-scale complex cartoon face images dataset with numerous variations. Thus, it is an ideal candidate dataset for feature extraction work and improves model performance on human face images. This thesis conducted three experiments on Simpson dataset. These experiments are focused on feature extraction module by substituting default feature extractor with VGG19, VGGFace2 and EfficientNet-B0-NoisyStudent (Xie et al., 2019) in SRGAN.

Additionally, several experimental changes are made to default SRGAN hyper-parameters. These changes are listed in table 5.1. Table 5.4 neatly consolidates and summarizes three experiments along with respective model architectures, hyper-parameters and pre-trained network weights for Simpson dataset.

Dataset name	Simpson dataset
Experiment 1: SRGAN-VGG19	Model architecture = Use VGG19 feature extractor  Pre-trained network weight = VGG19 Imagenet VGG19 Imagenet loss function = MSE for Accuracy metrics Adversarial loss = binary_crossentropy Epochs = 51, 101, 151, 201 and 401
Experiment 2: SRGAN-EfficientNetB0	Model architecture = Use EfficientNetB0 feature extractor  Pre-trained network weight = EfficientNetB0 Imagenet EfficientNetB0 Imagenet loss function = MSE for Accuracy metrics Adversarial loss = binary_crossentropy Epochs = 51, 101, 151, 201 and 401

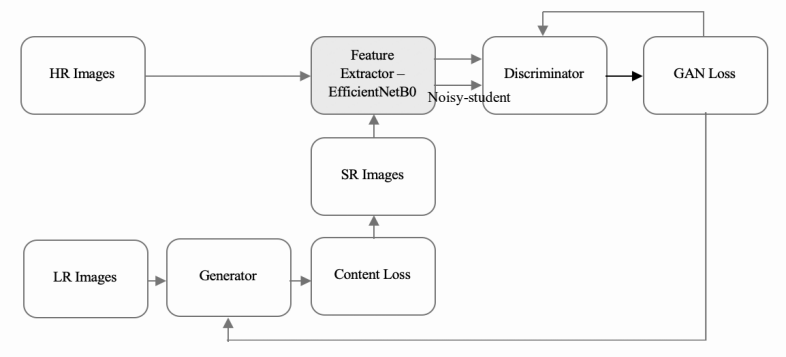
Experiment 3: SRGAN-EfficientNetB0-Noisy-student	Model architecture = Use EfficientNetB0-NoisyStudent feature extractor  Pre-trained network weight = EfficientNetB0 noisy-student EfficientNetB0 noisy-student loss function = MSE for Accuracy metrics Adversarial loss = binary_crossentropy Epochs = 51, 101, 151, 201 and 401

Table 5.4 Simpson dataset experiments details

Experiments listed in table 5.4 meet the objective and scope of this thesis stated in chapter 1, section 1.3 and chapter 1, section 1.4. In the next section, we deep dive into results by performing thorough interpretation of visualization obtained through experiments.

5.4.4.1 Generator and Discriminator loss analysis

This thesis conducted experiments which are focused on feature extraction module by substituting default feature extractor with VGG19, VGGFace2 and EfficientNetB0:Noisy-Student, along with tuning of default hyper-parameters in SRGAN. As per research conducted by (Xie et al., 2019), using EfficientNetB0:noisy-student pre-trained learning network improves classification performance of models. The figure 5.13 represents Generator and Discriminator loss values obtained after experiments.

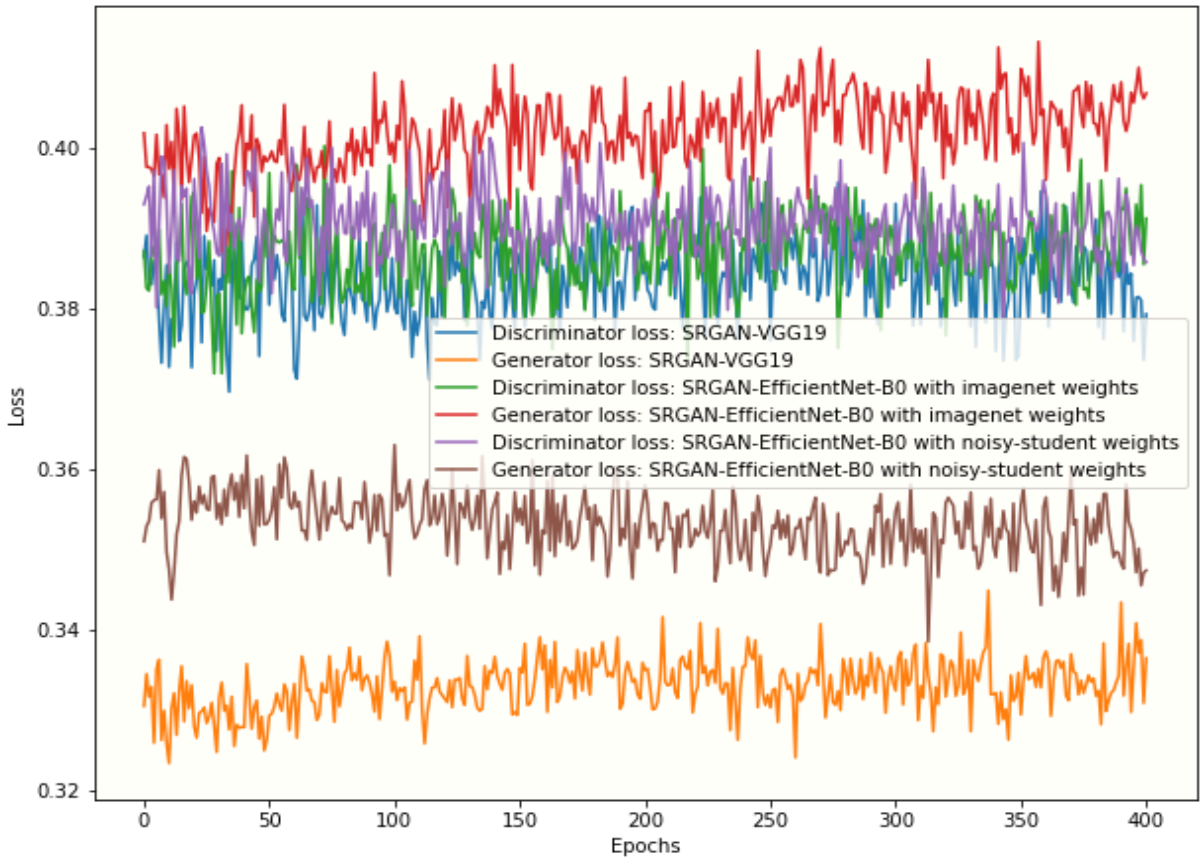


Figure5.13
Generator and Discriminator loss value analysis for Simpson dataset.

It is evident from figure 5.13 that Generator and Discriminator loss values on Simpson dataset is quite similar to loss values obtained from CelebA dataset for VGG19 pre-trained weights. For EfficientNetB0 and EfficientNetB0:noisy-student, Generator loss is significantly lower compared to loss values obtained from CelebA dataset. This becomes possible because of increased number of epochs, up-to 401.

We have stated this fact while evaluating CelebA and LFW loss results. Another evident outcome is lower Generator loss with EfficientNetB0:noisy-student pre-trained weights compared to EfficientNetB0 weights. The experimental changes in hyper-parameters and use of pre-trained network weights results in very stable and low (Brownlee, 2021) Generator loss and Discriminator loss for all pre-trained learning weights. These results are much lower and stable compared to standard SRGAN implementation.

5.4.4.2 PSNR and SSIM analysis

We have evaluated model performance on both metrics to maintain parity with industry practices. We skipped feature extraction for left eye and right eye because Simpson dataset is a collection of cartoon character face images with improper or missing facial features. PSNR and SSIM values are plotted for experiments listed in table 5.4.

- **Full images analysis:** It is evident from figure 5.14a and 5.14b that all pre-trained networks have steep upward trajectory for PSNR because of higher number of epochs and less complex features on cartoon character images. VGG19 performs better on SSIM metrics and shows an upward trajectory, in-line with increasing number of epochs.

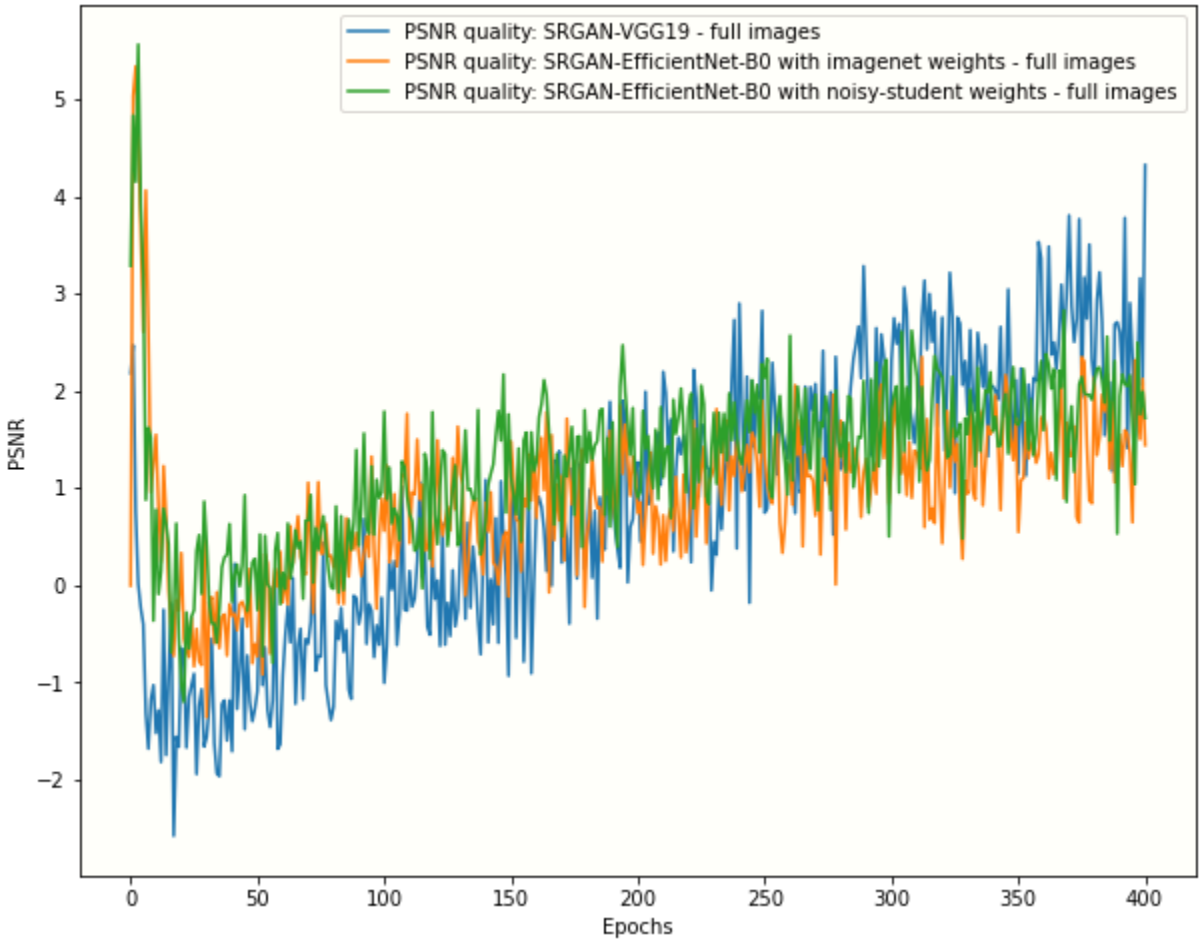


Figure5.14a
PSNR value analysis on full image for Simpson dataset.

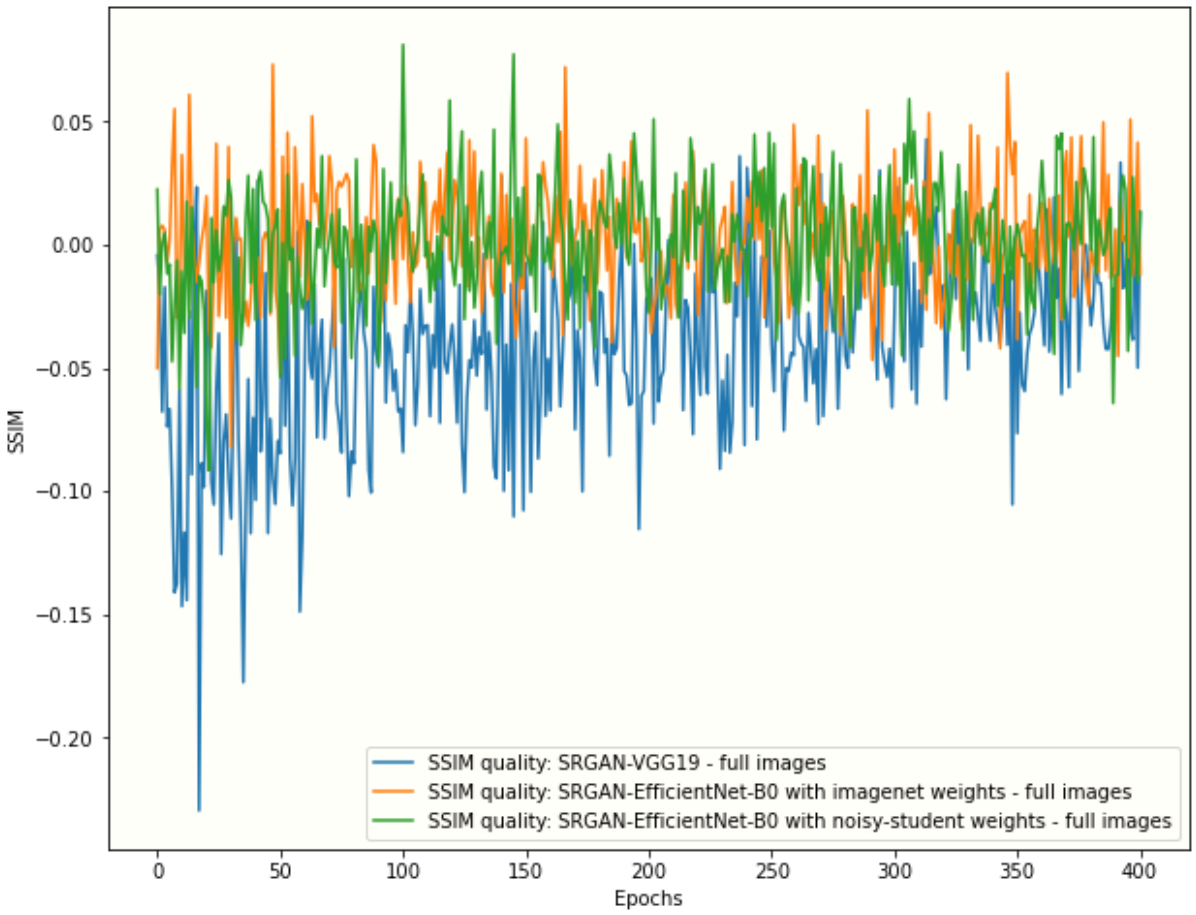


Figure5.14b
SSIM value analysis on full image for Simpson dataset.

- **Feature extraction analysis – face extraction:** It is evident from figure 5.15a and 5.15b that all pre-trained networks have upward trajectory for PSNR while VGG19 have steeper upward trajectory for PSNR values, in-line with increasing number of epochs. VGG19 performs better on SSIM metrics.

Figure5.15a
PSNR value analysis on extracted faces for Simpson dataset.

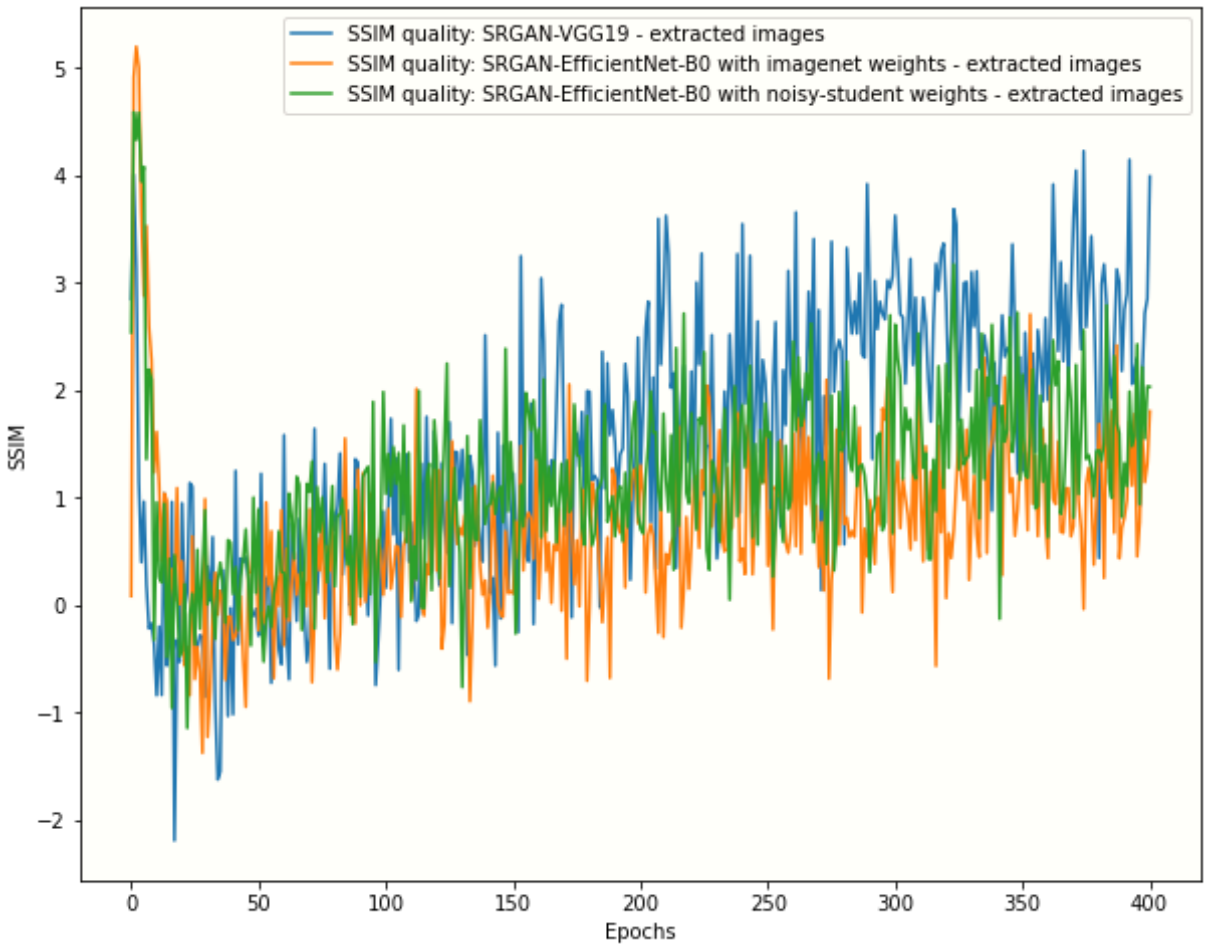


Figure5.15b
SSIM value analysis on extracted faces for Simpson dataset.

5.4.4.3 Simpson conclusion

Experiments conducted on Simpson dataset clearly shows that we are not only able to match but also surpass Generator loss i.e. produce lower loss for EfficientNetB0:noist-student pre-trained learning network backbones compared to CelebA EfficientNetB0 pre-trained learning network backbone. It reaffirms our objectives that selection of custom hyper-parameter values with pre-trained learning network backbone, along with relatively higher number of epochs have resulted in remarkably stable and low level of loss (Brownlee, 2021) for Generator and Discriminator compared to other implementations based on SRGAN (Ledig et al., 2017; Li et al., 2020; Zhang et

al., 2020). The use of pre-trained learning network backbone shows that by using an already learned network with appropriate hyper-parameter values, an increasingly high level of effectiveness can be maintained in feature extraction, even with sub 500 number of epochs. Out of three pre-trained weights, VGG19 shows great potential for feature extraction, in-line with increasing number of epochs. The PSNR and SSIM number could have been very high if epochs counts could be 10^5 or more, together with high batch size which couldn't be achieved given memory constraints on lab hardware.

5.5 Summary

We discussed implementation of proposed generative adversarial network models, evaluation metrics and explained results obtained from experiments conducted for feature extraction on three image-aligned datasets with pre-trained learning network weights. We used SRGAN as a base GAN model to perform experiments on three image aligned datasets for super resolution feature extraction.

For feature extraction, we substituted feature extractor in SRGAN by VGG19, VGGFace2 and EfficientNet-B0 pre-trained learning weights, together with various changes applied to hyper-parameters. To meet stated thesis objectives, more than nine experiments have been performed. The evaluation metrics used to measure performance of feature extraction are PSNR and SSIM. The evaluation metric used to measure loss function outcome for Generator and Discriminator is MSE loss with binary_crossentropy.

It is evident from results that experimental changes in hyper-parameters and use of pre-trained learning network weights resulted in very stable and low loss for Generator and Discriminator across all pre-trained weights. This is a great achievement considering epochs count used was between 200 to 500. This is remarkable compared to other implementations based on

SRGAN (Ledig et al., 2017; Li et al., 2020; Zhang et al., 2020) where epochs count is 10^5 or even more. Experiments conducted for feature extractions shows great results as well.

The use of pre-trained learning network backbone shows that by using an already learned network with appropriate hyper-parameter values, an increasingly high level of effectiveness can be maintained in feature extraction. Out of three pre-trained weights, mostly VGG19 and in some cases, EfficientNetB0 show great results for feature extraction. The PSNR and SSIM number could have been very high if epochs counts could be 10^5 or more, together with high batch size which couldn't be achieved given memory constraints on lab hardware.

These experiments met objectives and scope of this thesis stated in chapter 1, section 1.3 and chapter 1, section 1.4.

CHAPTER VI:

CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

This chapter concludes the thesis dissertation by providing a conclusive view of work accomplished, extensive literature review, model tuning, results and findings, and suggestions for future work. This dissertation experimented with Super Resolution Generative Adversarial Networks models for feature extraction work on diverse set of face images consolidated in image aligned large datasets. Various hyper-parameters were tuned and multiple pre-trained learning network weights were used in this work to evaluate Super Resolution Generative Adversarial Network performance in feature extraction work. This chapter summarizes findings obtained from experiments while also paving the road for improvements and future work.

6.2 Discussions and Conclusions

The main aim of this research was to study and hopefully improve State-of-the-Art in Super Resolution Generative Adversarial Networks performance for feature extraction on different large-scale complex face images datasets by performing various experiments. To fulfill this aim, multiple objectives were defined early at research proposal stage and these objectives were explored throughout this dissertation while keeping scope of study in check. To meet these objectives, these tasks were completed:

- This dissertation accomplished an extensive literature survey on Super Resolution, Convolutional Neural Network for Super Resolution, Deep Convolutional Neural Networks for Super Resolution and Generative Adversarial Networks for Super Resolution. The survey concluded with a comprehensive list of Super Resolution Generative Adversarial Networks methods, their performance measured on SSIM and PSNR benchmarks, pros and

cons of each method, real world uses examples and gaps in Generative Adversarial Network based Super Resolution experiments.

- Conducted extensive research on hyper-parameters to modify and tune specific parameters. This leads to creation of modified SRGAN method for feature extraction.
- Conducted comparative analysis with existing State-of-the-art work by using modified SRGAN method architecture with various pre-trained transfer learning backbones on multiple large-scale complex face images datasets for feature extraction from images.
- For feature extraction, we substituted feature extractor in SRGAN by VGG19, VGGFace2, EfficientNetB0 and EfficientNetB0:noisy-student pre-trained transfer learning weights, together with various changes applied to hyper-parameters.
- Evaluated the performance and effectiveness of feature extraction on industry standard SSIM and PSNR benchmarks, along with evaluation of Generator and Discriminator loss.

This dissertation used three datasets of large challenging face images which gave a very good opportunity to experiments with cross domain learning and improve model performance in real world. The large scale, quantity and variations of facial images created opportunities to tinker with model parameters. This resulted in improved loss values and feature extraction performance in fewer number of epochs.

The datasets used were image aligned versions of CelebA dataset, image aligned deep-funneled version of LFW dataset and image aligned version of Simpson dataset. The image aligned version of datasets eliminated dataset preparation activities. Thus, tremendously speeding up technical implementation work.

The findings proved that experimental changes in hyper-parameters and use of pre-trained learning network weights resulted in very stable and low loss for Generator and Discriminator across all pre-trained weights. This is a great achievement considering epochs count used was

between 200 to 500. This is remarkable compared to other State-of-the-art implementations based on SRGAN where epochs count is 10^5 or even more. The use of pre-trained learning network backbone illustrated that by using an already learned network with appropriate hyper-parameter values, an increasingly high level of effectiveness can be maintained in feature extraction. Out of three pre-trained weights, mostly VGG19 and in some cases, EfficientNetB0 show great results for feature extraction.

6.3 Contribution to knowledge

The field of computer vision is in constant need to better and faster architectures. Feature extraction space in Super Resolution Generative Adversarial Network is not different. The existing research papers are mostly focused on up-to 4x scaling and training for very small datasets. One of the aims of this research was to evaluate feature extraction performance on large-scale complex face images datasets when used with pre-trained network learning backbones to produce stable Generative Adversarial Networks architectures which uses fewer epochs while still providing great performance.

By completing an extensive survey on existing State-of-The-Art Super Resolution Generative Adversarial Networks and performing various experiments on feature extraction space, this study would help fellow researchers to further improve GAN models for SR in the field of feature extraction, improve perceptual visual quality of generated image at higher upscaling factors with a smaller number of epochs. This will result in faster and more efficient GAN models.

6.4 Future Recommendations

The field of computer vision is expanding rapidly in the space of feature extraction using Super Resolution Generative Adversarial Networks. After conducting an extensive survey on

current state of Super Resolution Generative Adversarial Network and performing various experiments, the future recommendation of this study would be:

- Conduct experiments to find out appropriate values of hyper-parameters which should result in better and faster SR GAN model architecture for feature extraction. Some examples are: Batch size, filter size, number of epochs and learning rates.
- Explore more pre-trained network weights. This study has shown that use of pre-trained network weights can improve effective performance of feature extraction.
- Conduct experiments on super large image datasets with very high degree of image variations.
- Conduct experiments on high up-sampling factor of 8x and more.
- Conduct experiments on various combination of loss functions, filters, position of layers and training strategies.
- Existing GAN models are Compute and Memory intensive. This was evident in experiments conducted in this study where compute power resources were a major factor to not use bigger batch size and more epochs. This area requires more exploration. This will help push GAN in real time high framerate video streaming, conversion and feature extraction.

APPENDIX A

PROGRAM CODE

Immense care has been taken in providing program code. Model's business and inter-component data communication logic have been abstracted in a way to protect individual work and intellectual property of author. The code below provides all necessary steps and options to test and validate thesis findings without making it possible to recreate exact replica of Model's underlying business logic and inter-component data communication logic. For collaboration, kindly get it touch with thesis author.

The code can be found at - https://github.com/codeisthelife/amdbassbmmarch2024

REFERENCES

alexattia, (2018) *The Simpsons Characters Data*. [online] Available at: https://github.com/alexattia/SimpsonRecognition.

Anwar, S., Khan, S. and Barnes, N., (2019) A Deep Journey into Super-resolution: A survey. [online] pp.1–21. Available at: http://arxiv.org/abs/1904.07523.

Arxiv-sanity, (n.d.) *Arxiv Sanity Preserver*. [online] Available at: http://www.arxiv-sanity.com/.

Bell-Kligler, S., Shocher, A. and Irani, M., (2019) Blind super-resolution kernel estimation using an internal-GAN. *Advances in Neural Information Processing Systems*, [online] 32788535, pp.1–10. Available at: https://arxiv.org/abs/1909.06581v6.

Bhatia, V. and Kumar, Y., (2020) Attaining Real-Time Super-Resolution for Microscopic Images Using GAN. [online] Available at: http://arxiv.org/abs/2010.04634.

Brownlee, J., (2021) How to Identify and Diagnose GAN Failure Modes. *Machine Learning Mastery*, [online] p.89. Available at: https://machinelearningmastery.com/practical-guide-to-gan-failure-modes/.

Bulat, A., Yang, J. and Tzimiropoulos, G., (2018) To learn image super-resolution, use a GAN to learn how to do image degradation first. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [online] pp.187–202. Available at: https://arxiv.org/pdf/1807.11458.pdf.

Cai, J., Han, H., Shan, S. and Chen, X., (2020) FCSR-GAN: Joint Face Completion and Super-Resolution via Multi-Task Learning. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, [online] 22, pp.109–121. Available at: https://ieeexplore.ieee.org/document/8890717/.

Cao, Q., Shen, L., Xie, W., Parkhi, O.M. and Zisserman, A., (2018) VGGFace2: A dataset for recognising faces across pose and age. In: *Proceedings - 13th IEEE International*

Conference on Automatic Face and Gesture Recognition, FG 2018. [online] IEEE, pp.67–74. Available at: https://ieeexplore.ieee.org/document/8373813/.

Cao, S., Wu, C. and Krähenbühl, P., (2020) Lossless Image Compression through Super-Resolution. [online] Available at: https://arxiv.org/pdf/2004.02872.pdf.

Cheon, M., Kim, J.H., Choi, J.H. and Lee, J.S., (2019) Generative adversarial network-based image super-resolution using perceptual content losses. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [online] pp.51–62. Available at: https://arxiv.org/abs/1809.04783v2.

Demiray, B.Z., Sit, M. and Demir, I., (2020) D-SRGAN: DEM Super-Resolution with Generative Adversarial Networks. [online] Available at: http://arxiv.org/abs/2004.04788.

Ditria, L., Meyer, B.J. and Drummond, T., (2020) OpenGAN: Open Set Generative Adversarial Networks. [online] pp.1–26. Available at: http://arxiv.org/abs/2003.08074.

Dong, C., Loy, C.C., He, K. and Tang, X., (2016) Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, [online] 382, pp.295–307. Available at: https://arxiv.org/pdf/1501.00092v3.pdf.

Fu, Y., Chen, W., Wang, H., Li, H., Lin, Y. and Wang, Z., (2020) AutoGAN-Distiller: Searching to Compress Generative Adversarial Networks. [online] Available at: http://arxiv.org/abs/2006.08198.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y., (2014) Generative adversarial nets. In: *Advances in Neural Information Processing Systems*. [online] pp.2672–2680. Available at: http://arxiv.org/abs/1406.2661.

Guo, Y., Chen, Q., Chen, J., Wu, Q., Shi, Q. and Tan, M., (2019) Auto-Embedding Generative Adversarial Networks for High Resolution Image Synthesis. *IEEE Transactions on*

Multimedia, [online] 2111, pp.2726–2737. Available at: https://arxiv.org/abs/1903.11250v2.

He, K., Zhang, X., Ren, S. and Sun, J., (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. [online] IEEE, pp.770–778. Available at: http://ieeexplore.ieee.org/document/7780459/.

Horé, A. and Ziou, D., (2010) Image quality metrics: PSNR vs. SSIM. *Proceedings - International Conference on Pattern Recognition*, [online] pp.2366–2369. Available at: https://ieeexplore.ieee.org/document/5596999.

Hu, X., Liu, X., Wang, Z., Li, X., Peng, W. and Cheng, G., (2019) RTSRGAN: Real-time super-resolution generative adversarial networks. In: *Proceedings - 2019 7th International Conference on Advanced Cloud and Big Data, CBD 2019*. [online] IEEE, pp.321–326. Available at: https://ieeexplore.ieee.org/document/8916480/.

Huang, R., Zhang, S., Li, T. and He, R., (2017) Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis. In: *Proceedings of the IEEE International Conference on Computer Vision*. [online] IEEE, pp.2458–2467. Available at: https://arxiv.org/pdf/1704.04086.pdf.

IEEE, (2002) *IEEE Xplore*. [online] IEEE Transactions on Education. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6348063%0Ahttps://ieeexplore-ieee-org.ezproxy-eres.up.edu/stamp/stamp.jsp?arnumber=1013875%0Ahttps://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8936407%0Ahttps://ieeexplore.ieee.org/stamp/stam.

Indradi, S.D., Arifianto, A. and Ramadhani, K.N., (2019) Face Image Super-Resolution Using Inception Residual Network and GAN Framework. In: *2019 7th International Conference*

on Information and Communication Technology (ICoICT). [online] IEEE, pp.1–6. Available at: https://ieeexplore.ieee.org/document/8835253/.

Jessica Li, (2018a) *CelebFaces Attributes (CelebA) Dataset | Kaggle*. [online] Available at: https://www.kaggle.com/jessicali9530/celeba-dataset.

Jessica Li, (2018b) *Labelled Faces in the Wild (LFW) Dataset*. [online] Available at: https://www.kaggle.com/jessicali9530/lfw-dataset.

Johnson, J., Alahi, A. and Fei-Fei, L., (2016) Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [online] pp.694–711. Available at: https://arxiv.org/pdf/1603.08155.pdf.

Keras.io, (2020) *Keras: the Python deep learning API*. [online] Available at: https://keras.io/.

Kim, J., Lee, J.K. and Lee, K.M., (2016) Accurate image super-resolution using very deep convolutional networks. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, [online] 2016-Decem, pp.1646–1654. Available at: https://arxiv.org/abs/1511.04587v2.

Kim, S.Y., Oh, J. and Kim, M., (2019) JSI-GAN: GAN-Based Joint Super-Resolution and Inverse Tone-Mapping with Pixel-Wise Task-Specific Filters for UHD HDR Video. [online] Available at: http://arxiv.org/abs/1909.04391.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. and Shi, W., (2017) Photo-realistic single image super-resolution using a generative adversarial network. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, [online] 2017-Janua, pp.105–114. Available at: https://arxiv.org/pdf/1609.04802.pdf.

Lfw, (2013) *Labeled Faces in the Wild: Results*. [online] Available at:

http://rd.springer.com/chapter/10.1007/978-3-319-25958-1_8%0Ahttp://vis-www.cs.umass.edu/lfw/results.html.

Li, J., Zhou, Y., Ding, J., Chen, C. and Yang, X., (2020) ID Preserving Face Super-Resolution Generative Adversarial Networks. *IEEE Access*, [online] 8, pp.138373–138381. Available at: https://ieeexplore.ieee.org/document/9146819/.

Liu, Z., Luo, P., Wang, X. and Tang, X., (2018) *Large-scale celebfaces attributes (celeba) dataset*. [online] Retrieved August. Available at: http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.

Ma, C., Rao, Y., Cheng, Y., Chen, C., Lu, J. and Zhou, J., (2020) Structure-Preserving Super Resolution With Gradient Guidance. [online] pp.7766–7775. Available at: https://arxiv.org/abs/2003.13081v1.

Mahapatra, D., Bozorgtabar, B. and Garnavi, R., (2019) Image super-resolution using progressive generative adversarial networks for medical image analysis. *Computerized Medical Imaging and Graphics*, [online] 71, pp.30–39. Available at: https://arxiv.org/abs/1902.02144v2.

OpenCV, (n.d.) *OpenCV*. [online] Available at: https://opencv.org/.

PaperWithCode, (n.d.) *Papers with Code : the latest in machine learning*. [online] Available at: https://paperswithcode.com.

Pathak, H.N., Li, X., Minaee, S. and Cowan, B., (2019) Efficient Super Resolution for Large-Scale Images Using Attentional GAN. *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, [online] pp.1777–1786. Available at: https://arxiv.org/abs/1812.04821v4.

Rakotonirina, N.C. and Rasoanaivo, A., (2020) ESRGAN+ : Further Improving Enhanced Super-Resolution Generative Adversarial Network. *ICASSP, IEEE International Conference on*

Acoustics, Speech and Signal Processing - Proceedings, [online] 2020-May, pp.3637–3641. Available at: https://arxiv.org/abs/2001.08073v2.

Simonyan, K. and Zisserman, A., (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, [online] pp.1–14. Available at: https://arxiv.org/pdf/1409.1556.pdf.

Singh, N.K. and Raza, K., (2020) Medical Image Generation using Generative Adversarial Networks. [online] pp.1–19. Available at: https://arxiv.org/abs/2005.10687v1.

Sood, R., Topiwala, B., Choutagunta, K., Sood, R. and Rusu, M., (2019) An Application of Generative Adversarial Networks for Super Resolution Medical Imaging. *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, [online] pp.326–331. Available at: https://arxiv.org/abs/1912.09507v1.

Tan, M. and Le, Q. V., (2019) EfficientNet: Rethinking model scaling for convolutional neural networks. *36th International Conference on Machine Learning, ICML 2019*, [online] 2019-June, pp.10691–10700. Available at: https://arxiv.org/pdf/1905.11946v5.pdf.

TensorFlow, (n.d.) *TensorFlow*. [online] Available at: https://www.tensorflow.org/.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C.C., Qiao, Y. and Tang, X., (2018a) ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. [online] Available at: http://arxiv.org/abs/1809.00219.

Wang, Y., Perazzi, F., McWilliams, B., Sorkine-Hornung, A., Sorkine-Hornung, O. and Schroers, C., (2018b) A Fully Progressive Approach to Single-Image Super-Resolution. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. [online] IEEE, pp.977–97709. Available at: https://arxiv.org/abs/1804.02900.

Wu, B., Duan, H., Liu, Z. and Sun, G., (2017) SRPGAN: Perceptual Generative Adversarial

Network for Single Image Super Resolution. [online] Available at: http://arxiv.org/abs/1712.05927.

Xie, Q., Luong, M., Hovy, E. and Le, Q. V, (2019) Self-training with Noisy Student improves ImageNet classification. [online] Available at: http://arxiv.org/abs/1911.04252.

Yang, S., Wang, H., Wang, Y., Li, J. and Wang, Y., (2020) Super-resolution reconstruction algorithm based on deep learning mechanism and wavelet fusion. *Beijing Hangkong Hangtian Daxue Xuebao/Journal of Beijing University of Aeronautics and Astronautics*, [online] 461, pp.189–197. Available at: https://arxiv.org/pdf/1907.10213v1.pdf.

Yuan, Y., Liu, S., Zhang, J., Zhang, Y., Dong, C. and Lin, L., (2018) Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. [online] pp.814–823. Available at: https://arxiv.org/pdf/1809.00437v1.pdf.

Zhang, G., Kan, M., Shan, S. and Chen, X., (2018) Generative Adversarial Network with Spatial Attention for Face Attribute Editing. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. [online] pp.422–437. Available at: http://link.springer.com/10.1007/978-3-030-01231-1_26.

Zhang, K., Zhang, Z., Li, Z. and Qiao, Y., (2016) Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, [online] 2310, pp.1499–1503. Available at: https://arxiv.org/pdf/1604.02878.pdf.

Zhang, Y., Zheng, Z. and Hu, R., (2020) Super Resolution Using Segmentation-Prior Self-Attention Generative Adversarial Network. [online] Available at: http://arxiv.org/abs/2003.03489.

Zhong, S. and Zhou, S., (2020) Optimizing Generative Adversarial Networks for Image Super Resolution via Latent Space Regularization. [online] Available at:

http://arxiv.org/abs/2001.08126.

Zhong, X., Qu, X. and Chen, C., (2019) High-quality face image super-resolution based on Generative Adversarial Networks. In: *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*. [online] IEEE, pp.1178–1182. Available at: https://ieeexplore.ieee.org/document/8998075.