Deep Learning for Robust Super-resolution

A thesis submitted to the
College of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Electrical and Computer Engineering
University of Saskatchewan
Saskatoon, Canada

By
Mahdiyar Molahasani Majdabadi

Permission to Use

In presenting this thesis/dissertation in partial fulfillment of the requirements for a Post-graduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis/dissertation in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis/dissertation work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis/dissertation or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis/dissertation.

Disclaimer

Reference in this thesis to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan, and shall not be used for advertising or product endorsement purposes

Request for permission to copy or to make any other use of material in this thesis in whole or in part should be addressed to:

Head of the Department of Electrical and Computer Engineering
57 Campus Drive
University of Saskatchewan
Saskatoon, Saskatchewan, Canada
S7N 5A9

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9 Canada

Abstract

Super Resolution (SR) is a process in which a high-resolution counterpart of an image is reconstructed from its low-resolution sample. Generative Adversarial Networks (GAN), known for their ability of hyper-realistic image generation, demonstrate promising results in performing SR task. High-scale SR, where the super-resolved image is notably larger than low-resolution input, is a challenging but very beneficial task. By employing an SR model, the data can be compressed, more details can be extracted from cheap sensors and cameras, and the noise level will be reduced dramatically. As a result, the high-scale SR model can contribute significantly to face-related tasks, such as identification, face detection, and surveillance systems. Moreover, the resolution of medical scans will be notably increased. So more details can be detected and the early-stage diagnosis will be possible for many diseases such as cancer. Moreover, cheaper and more available scanning devices can be used for accurate abnormality detection. As a result, more lives can be saved because of the enhancement of the accuracy and the availability of scans.

In this thesis, the first multi-scale gradient capsule GAN for SR is proposed. First, this model is trained on CelebA dataset for face SR. The performance of the proposed model is compared with state-of-the-art works and its supremacy in all similarity metrics is demonstrated. A new perceptual similarity index is introduced as well and the proposed architecture outperforms related works in this metric with a notable margin. A robustness test is conducted and the drop in similarity metrics is investigated. As a result, the proposed SR model is not only more accurate but also more robust than the state-of-the-art works.

Since the proposed model is considered as a general SR system, it is also employed for prostate MRI SR. Prostate cancer is a very common disease among adult men. One in seven Canadian men is diagnosed with this cancer in their lifetime. SR can facilitate early diagnosis and potentially save many lives. The proposed model is trained on the Prostate-Diagnosis and PROSTATEx datasets. The proposed model outperformed SRGAN, the state-of-the-art prostate SR model. A new task-specific similarity assessment is introduced as well. A classifier is trained for severe cancer detection and the drop in the accuracy of this model

when dealing with super-resolved images is used for evaluating the ability of medical detail reconstruction of the SR models. This proposed SR model is a step towards an efficient and accurate general SR platform.

Acknowledgments

I would like to thank my supervisor, Professor Seok-Bum Ko for his dedicated support and guidance throughout my M.Sc program. Without his encouragement and assistance, this dissertation would not have been possible. I also thank the following people who have helped me undertake this research: My lab-mates, especially Arman Haghanifar for his innovative ideas, my B.Sc. supervisor, Dr. B. Shokouhi for his support and encouragement, Dr. S. Deivalakshmi and her team from the National Institute of Technology, Tiruchirappalli for her help and support in taking this thesis to the next level, and my family, who have paved this road for me with their endless support and patience. My sincere thanks to the Department of Electrical and Computer Engineering, University of Saskatchewan.

To the one who brought light and joy to my life.

Table of Contents

List of Abbreviations

AI	Artifical Inteligence
CapsNet	Capsule Network
CLAHE	Contrast Limited Adaptive Histogram Equalization
CNN	Convolutional Neural Network
DRE	Digital Rectal Exam
FAN	Face Alignment Network
FSIM	Feature SIMilarity
GAN	Generative Adverserial Network
HR	High-Resolution
LOF	List of Figures
LOT	List of Tables
LR	Low-Resolution
LSTM	Long Short-Term Memory
MLP	Multi-Layer Perceptron
MOS	Mean Opinion Score
MPNR	Mirror-Patch-based Neighbor Representation
MRF	Markov random field
MS-SSIM	Multi-Scale Structural SIMilarity
MSE	Mean Square Error
MSG-CapsGAN	Multi-Scale Gradient Capsule GAN

PSA	Prostate-Specific Antigen
PSNR	Peak Signal-to-Noise Ratio
ReLU	Rectified Linear Unit
ReLU	Rectified linear unit
RNN	Recurrent Neural Networks
RRDB	Residual-in-Residual Dense Block
SR	Super-Resolution
SSIM	Structural SIMilarity
TSSA	Task-Specific Similarity Assessment
VAE	Variational Auto-Encoder
WL	Windows Level
WW	Windows Width

List of Tables

List of Figures

1. Introduction

Super-Resolution (SR) methods can increase the number of pixels and the quality of an image. As a result, a small noisy blurry picture can become a clear large image. In this thesis, a novel general SR system is proposed. The latest advancements and trends in deep learning and SR is presented in this chapter, followed by a summary of contributions of this thesis and the publications and submissions during the M.Sc. program.

1.1 Research problem and objectives

SR is is an ill-posed problem. The reason is the fact that different High-Resolution (HR) images can have the same Low-Resolution (LR) counterpart. If the differences between the two images are small, then after down-sampling, the images will be identical. Hence, finding the most accurate reconstructed HR image is extremely challenging and impossible in some cases. Moreover, due to the lack of information in the LR input, filling the gaps in the super-resolved image needs prior knowledge. For example, in a 16×16 image, there are 256 pixels. Each pixel is represented by 3 8-bit values. In total, this image carries 3840 bits of information. However, after $\times 8$ SR, the reconstructed image will have 245760 bits of information. So, based on the available data in the LR image and the domain-specific prior knowledge of the model, the new 241920 bits should be determined.

These two major problems make accurate SR a very challenging task. Another important quality that should be present in the super-resolved image, other than the accuracy, is being realistic. Some methods, such as interpolations, can provide relatively accurate images but they look blurry and can easily be identified as fake images. Since the similarity of images is usually compared based on the pixel values, the models are forced to generate outputs

similar to the ground truth. However, two images with the same similarity (i.e. PSNR) can have a different level of being realistic. So it is very important to consider this in designing the model. The new pixel values should be determined in a way that the super-resolved image not only be accurate but also look photo-realistic. Based on the mentioned problems, it is very important to design a model which could generate accurate and visually pleasant HR images from LR inputs.

The latest advancements in the field of deep learning are employed in this thesis with a creative approach to form a novel architecture that could answer the need for a powerful general SR system. The performance of this system is analyzed to demonstrate its merits from different perspectives. The main objective of the SR system is accuracy. The reconstructed HR images should be loyal to ground truth. To evaluate this quality of the proposed system, various metrics are used. Because of the lack of proper metrics for SR system performance from a perspective of a deep learning model, a novel similarity metric is proposed. Moreover, to investigate the performance of the system in preserving medical details crucial for accurate diagnosis, a new task-specific similarity assessment is introduced, as well.

Another important objective of this work is to propose a general SR system that could perform well in various domains. In the literature, many SR systems are proposed for specific domains. These models can perform with a high performance only in one type of data. For example, many SR systems are designed specifically for face SR. The reason behind this is these models are developed based on a specific type of data architecture-wise. Moreover, the domain-specific information is utilized as well. This domain-specific information can be acquired in different ways such as labeled data and pre-trained feature extractors. For example, in the face SR, the model can have access to the facial attribute labels like gender, age, etc., or a classifier trained to find these labels can be embedded in the model and provide this information. In this thesis, in order to develop a general SR model, no domain-specific information or feature extraction model is used in the proposed architecture. So the model could be trained on any type of data. Furthermore, the robustness of the model became one of our highest priorities. Therefore, one of the newest deep learning models is used in our model, and the performance of the proposed system while facing transformation attack is

compared with the state-of-the-art works.

1.2 Motivations

Since the introduction of the Generative Adversarial Network (GAN), generative Deep Learning (DL) models flourish unprecedentedly. This new concept of training enables the DL model to synthesize hyper-realistic images. As a result, a great opportunity has been provided for the generative models to improve their performance significantly. One of the most popular generative tasks is SR where the high-resolution counterpart of a low-quality image is reconstructed. These powerful GAN-based models alongside another novel architecture called capsule network to motivate us to propose a new capsule GAN model for image SR. The proposed model can contribute to many face-related tasks and also can potentially save lives by providing the radiologist with higher quality medical scans.

1.2.1 Trends in Super-resolution

SR is a challenging ill-posed problem attempting to reconstructing the HR image from its LR counterpart. LR image can be modeled as follows [2].

$$x_{LR} = (x_{HR} \otimes K) \downarrow_s + n \tag{1.1}$$

Where x is the image, K is a blurry filter, $\downarrow_s$ is image down-sampling operation, and n is the additive noise. Fig. 1.1 illustrates a diagram of down-sampling, SR, and similarity assessment.

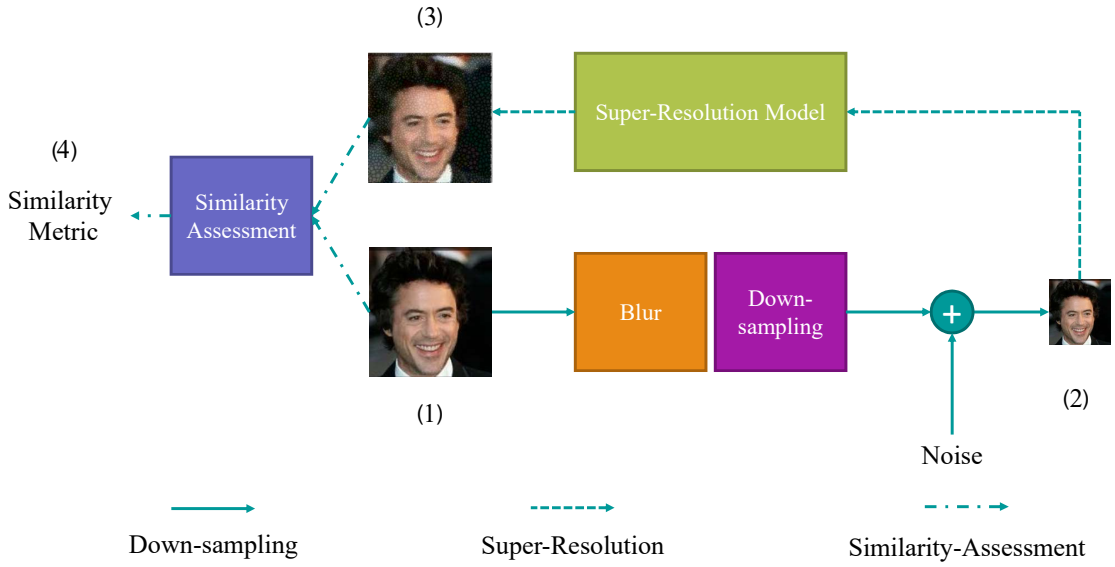


Figure 1.1: Super-resolution process with four steps: (1) ground-truth, (2) down-sampled image, (3) super-resolved sample, and (4) similarity metric.

In this thesis, the down-sampling block is implemented by an average function. A window of pixels is selected and the average of the pixel values is calculated. This value creates a pixel value in the LR image. This process is performed in each channel. Regarding the noise, it should reflect the characteristics of the noise of the sensor that the SR is going to implemented on its output. In this thesis, white Gaussian noise and salt and pepper noise are added to the image. The super-resolution model and the similarity metrics will be explained in the next chapter. Interpolation methods are very fast SR methods. These mathematical non-learning algorithms perform well in low-scale SR. However, they suffer from low-quality outputs and high computational cost when the SR scale increases. Bicubic interpolation and Lanczos resampling are two popular examples of these algorithms [10, 11].

More advanced and efficient SR methods are learning-based or so-called example-based algorithms. Most of these methods are usually using machine learning so they could learn from thousands of examples. In these approaches, an HR image is reconstructed by learning a mapping between LR and HR images. Fig. 1.2 compares the interpolation methods and Artificial Intelligent (AI) for SR.

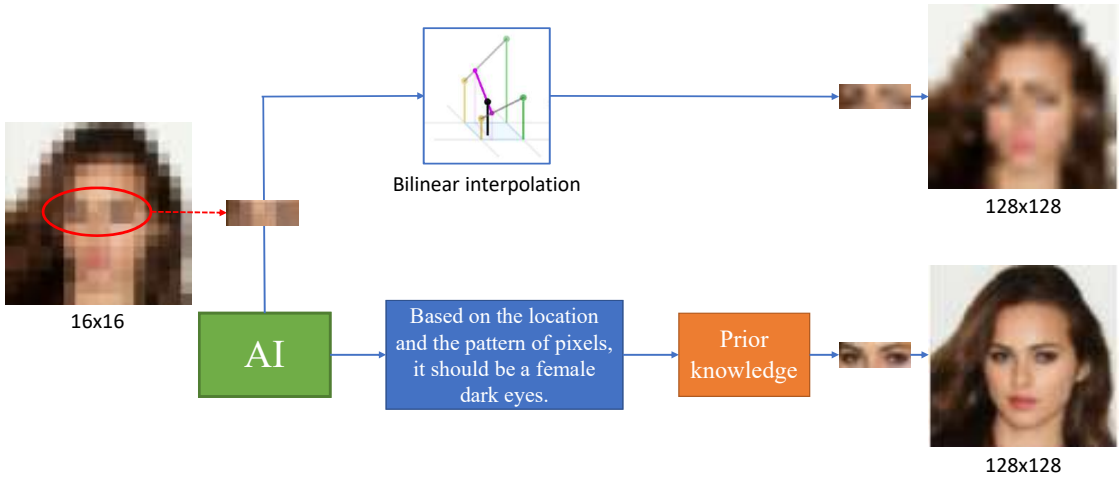


Figure 1.2: Comparison between bicubic interpolation [1] and AI based SR.

The interpolation methods are mathematical approaches aiming at finding new pixel values based on the neighbor values. However, AI-based models fill the gaps based on the content and the prior knowledge of the system acquired from the learning process. Basic machine learning algorithms has been widely used for SR, such as Markov random field (MRF) [12], Neighbor embedding methods [13], sparse coding methods [14], and random forest [15].

Since the introduction of deep learning, it has been widely used for SR [16–23]. These models achieved promising performance relying on their ability to learn high-level and complex features due to their complexity. The number of trainable parameters in these models varies from 20K to 43M [2]. More details on the deep learning architectures utilized in SR is provided in sections 2.1 and 3.1.2.

An important and challenging step in the SR task is image quality and similarity assessment, as depicted in Fig. 1.1. Many attempts have been made to propose task-specific image quality and/or similarity metrics. However, in most cases, Mean Square Error (MSE) or MSE based metrics such as Peak Signal-to-Noise Ratio (PSNR) are the main quality assessment tools. MSE has shown inadequate precision in reflecting the true quality of an image [2]. More recently, content-based metrics such as perceptual loss and Feature SIMilarity (FSIM) have been proposed [24]. It is an absolute necessity to define clear task-specific assessments

for the sake of fair comparison and optimization. In other words, having a metric that can reflect the true quality or similarity of images can facilitate model comparison and set a true goal for the machine learning model optimization. In order to peruse this goal, two novel similarity metrics are introduced in this thesis in section.

1.2.2 Trends in Deep Learning Architectures

Deep learning is a sub-category of machine learning. Unlike traditional learning-based methods, no handcrafted features are utilized in deep learning models. The informative and optimized features required for each task and dataset are established through the training process. Due to the complexity of these models, they can combine basic features to create higher-level features. Hence, all three steps of basic feature detection, hierarchical high-level feature generation, and decision making are performed simultaneously in a single training process [25].

Since the introduction of the Multi-Layer Perceptron (MLP) algorithm in 1960 [26], Artificial Neural Network (ANN) has become more popular for deep learning implementation because of two reasons, high capacity and hierarchical structure. In the 80's, the backpropagation algorithm was introduced and utilized for training MLP [27]. The next huge leap in the deep learning world was the introduction of Convolutional Neural Network (CNN) in 1989 [28]. Recurrent Neural Networks (RNN) were the next deep learning model proposed in 1990. These models can show temporal dynamic behavior [29]. Capsule network (CapsNet), is one of the latest advancements in deep learning, achieved the state-of-the-art result on classification problems [30]. Unlike CNN, CapsNet can learn the geometric relationship between features, hence, it is more robust. More details on CapsNet is presented in section 3.1.1. Fig. 1.3 compares the four mentioned types of deep learning layers.

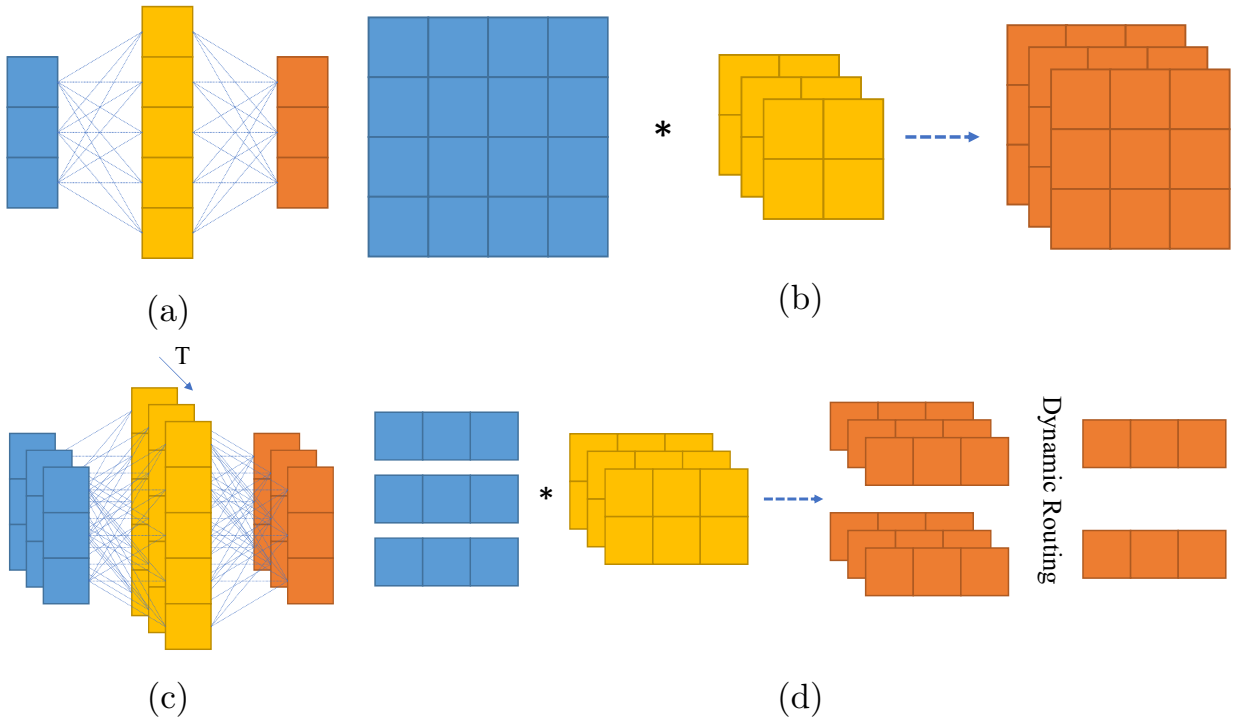


Figure 1.3: The architecture of (a) MLP, (b) CNN, (c) RNN, and (d) CapsNet.

The blue neurons represent the input layer, the weights are in yellow and the orange neurons are the layer's output. In MLPs, each neuron is connected to all neurons in the next layer, through a weight, called the synapse. Each neuron simply aggregated all weighted inputs and apply an activation function on it to form its output. The problem with MLPs is the high number of parameters. CNNs convolve several filters to the input tensor. Each filter represents a certain feature and it is trainable. The output of the convolution is a feature map on the input. The next convolutional layer combines the lower-level features extracted by the previous layer to detect high-level and more complex features. Each unit in a traditional RNN such as Long Short-Term Memory (LSTM), RNN has two inputs. First is the input of time step T and the second is the output of the unit in time step $T-1$. By connecting the output of each unit to its input, these units can benefit from short-term memory and it makes them the perfect choice for learning from time series or language-related tasks. CapsNet represents each feature with a vector. The values in the vector correspond to the geometrical pose of the feature and the length represents the probability

of its presence. Then. each feature makes a prediction for each higher-level feature. Finally, dynamic routing finds the agreement between the predictions and creates the final vector corresponding to each high-level feature or class.

As mentioned earlier, each neuron applies a non-linear function called activation to the weighted inputs. Many activation functions have been used in DL models so far. Fig. 1.4 demonstrates some of the most common activation functions.

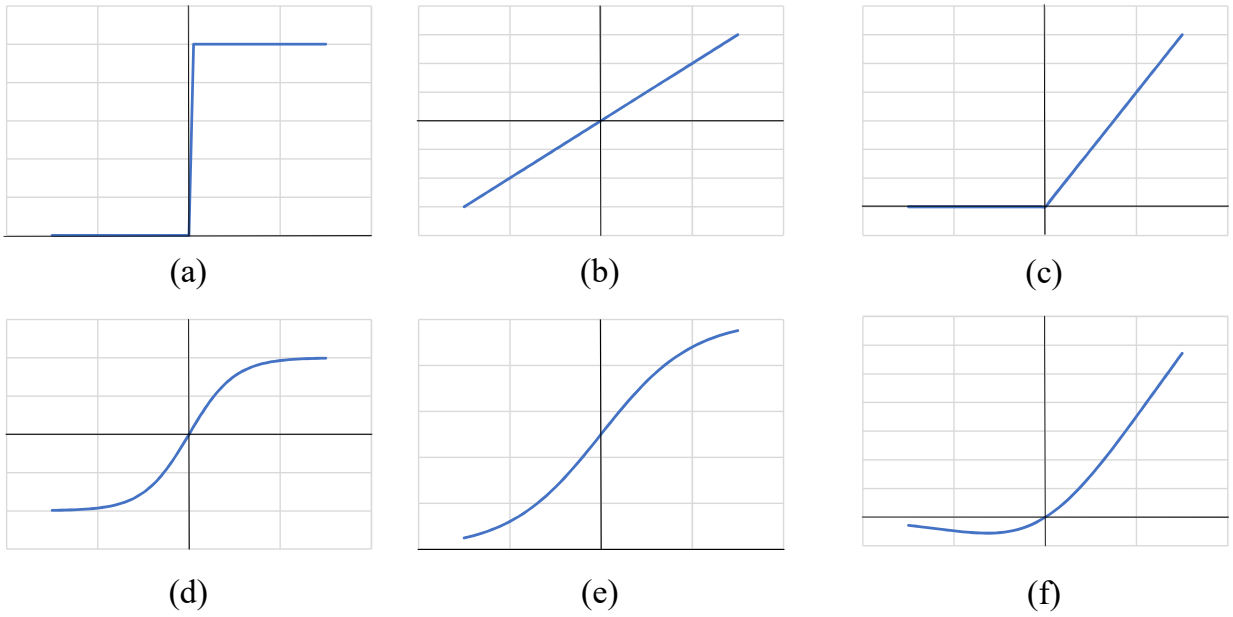


Figure 1.4: Some of the most common activation functions: (a) binary step, (b) identity, (c) Rectified Linear Unit (ReLU), (d) tanh, (e) sigmoid, and (f) swish.

The activation function is usually chosen based on its range of output and the architecture. For example, if the labels in a classification task are 0 and 1, the sigmoid function can be used. In the traditional LSTM, sigmoid and tanh are used and in CNNs, ReLU is one of the first choices.

In 2014, a new class of unsupervised deep learning models have been introduced [31]. By using the competitive learning paradigm, GANs have achieved unprecedented results in hyper-realistic sample generation. In section 3.1.2, the architecture and applications of this model are reviewed.

1.3 Super-resolution with Deep Learning

Face SR is a fast-growing field that aims to enhance the resolution of facial images. These systems attempt to reconstruct High-Resolution (HR) face images from their Low-Resolution (LR) counterparts accurately. Due to the importance of facial details on human perception, it is vital to preserving these facial details [32]. Face hallucination has widespread and crucial applications in various face-related systems such as face recognition, video surveillance system, and image editing [33].

To reconstruct HR images accurately, several challenges should be overcome. First, for large-scale face SR, reconstructing an accurate HR image is an arduous task due to the lack of information in the LR input. Second, it is required that the HR image not only possesses similarity to the ground truth but also has a photo-realistic appearance and seems natural. Finally, faces can appear in unlimited different poses. Hence, the facial SR system should be pose-invariant to generalize for various situations.

There are two categories of learning-based SR systems, local patch-based methods and global methods. In the first category, the system is trained to reconstruct a patch of an image at a time. Rajput and Arya propose a Mirror-Patch-based Neighbor Representation (MPNR) system for face hallucination [34]. Their findings suggest that this system is capable of filling missing parts in the face image, as well as super resolving noisy inputs. Several studies have proposed global methods in order to implement a precise face SR system. In these approaches, an HR image is reconstructed by learning a mapping between LR and HR images [18, 35–37]. More recently, deep learning is widely used for facial SR. Yu *et al.* propose a discriminative Generative Network to perform SR on aligned face images [38]. In their next works [39, 40], multiple spatial transformations are utilized in order to enhance the performance of the network. A cascade bi-network is introduced by Zhang *et al.* for reconstructing HR unaligned faces from very low-resolution inputs [41]. These deep learning-based methods improve the accuracy of face SR significantly. However, when the resolution of the input is fairly low, the performance of these networks reduces due to the distortion in the HR image.

To overcome the image distortion issue, different approaches have been proposed. Generative Adversarial Networks (GAN) have shown promising results in image synthesis [31, 42]. Hence, it has been widely used for the SR task due to its natural-looking outputs. Ledig *et al.* uses GAN for single image SR [43]. The perceptual loss function, which consists of adversarial loss and content loss, is used for training the GAN for the SR task. Another approach is using prior knowledge and attribute domain information for SR, especially face SR. This method is used for both Convolutional Neural Networks (CNN) and GANs. Kalarot *et al.* propose CAGFace, a fully convolutional patch-based face SR system. A component network is used to extract facial components and segment them. $4\times$ face SR is performed in their work. Lee *et al.* uses the information of the attribute domain as well as the image domain to reconstruct facial details in the HR image precisely [33]. Cheng *et al.* use a face shape estimation network for precise geometry estimation [3]. Kim *et al.* proposes a Face Alignment Network (FAN) for landmark heat map extraction. A new facial attention loss is applied for the training process based on their state-of-the-art FAN [4].

1.4 Contributions of the Thesis

In [44], we have proposed a novel Multi-Scale Gradient Capsule GAN (MSG-CapsGAN) for face SR to address the mentioned three challenges without using any attribute domain information. Capsule GAN and Multi-Scale Gradient GAN have been used for the SR task for the first time. The model has been trained on the CelebA dataset to increase the resolution of images from 16×16 to 128×128. This network has surpassed the state-of-the-art SR system in terms of PSNR.

In the next step and in [24], we have improved and redesigned the MSG-CapsGAN to enhance its performance and surpass previously introduced networks in all metrics. A robust capsule GAN is proposed with a novel residual transfer-learning-based generator for multi-scale face SR. The network is trained with an end-to-end training process without using any attribute domain information.

The proposed design has been optimized for prostate MRI SR. CheXNet has been embedded in the model for feature extraction. A new task-specific approach for image similarity

assessment is also introduced. The proposed model outperforms the state-of-the-art prostate SR system.

The contributions of this thesis are summarized as follows:

1. We utilize Capsule GAN and Multi-scale GAN for the SR task for the first time.
2. The proposed SR system surpassed the state-of-the-art systems in terms of PSNR, Structural SIMilarity (SSIM), Multi-Scale Structural SIMilarity (MS-SSIM), and Feature SIMilarity (FSIM).
3. The robustness of the network is evaluated and outperforms the state-of-the-art face SR system.
4. The model is used for prostate SR and outperforms the state-of-the-art prostate SR system.
5. A new task-specific metric for SR performance evaluation is introduced, called TSSA.

1.5 Publications and Submissions During M.Sc. Study

1.5.1 Published Journal

1. Molahasani Majdabadi, Mahdiyar, and Seok-Bum Ko. "Capsule GAN for robust face super resolution." Multimedia Tools and Applications 79, no. 41 (2020): 31205-31218. DOI: 10.1007/s11042-020-09489-y

   A Major portion of this paper is included in Chapter 3: Capsule GAN for RobustFace SR

2. Molahasani Majdabadi, Mahdiyar, Shahriar B. Shokouhi, and Seok-Bum Ko. "Efficient Hybrid CMOS/Memristor Implementation of Bidirectional Associative Memory Using Passive Weight Array." Microelectronics Journal 98 (2020): 104725. DOI: 10.1016/j.mejo.2020.104725

1.5.2 Published Conference

1. Molahasani Majdabadi, Mahdiyar, and Seok-Bum Ko. "Msg-capsgan: Multi-scale gradient capsule gan for face super resolution." In 2020 International Conference on Electronics, Information, and Communication (ICEIC), pp. 1-3. IEEE, 2020. DOI: 10.1109/ICEIC49074.2020.9051244

   A Major portion of this paper is included in Chapter 3: Capsule GAN for RobustFace SR

2. Haghanifar, Arman, Mahdiyar Molahasani Majdabadi, and Seok-Bum Ko. "Automated Teeth Extraction from Dental Panoramic X-Ray Images using Genetic Algorithm." In 2020 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1-5. IEEE, 2020. DOI: 10.1109/ISCAS45731.2020.9180937

1.5.3 Preprints

1. Haghanifar, Arman, Mahdiyar Molahasani Majdabadi, and Seokbum Ko. "Covid-cxnet: Detecting covid-19 in frontal chest x-ray images using deep learning." arXiv preprint arXiv:2006.13807 (2020).

2. Haghanifar, Arman, Mahdiyar Molahasani Majdabadi, and Seok-Bum Ko. "PaXNet: Dental Caries Detection in Panoramic X-ray using Ensemble Transfer Learning and Capsule Classifier." arXiv preprint arXiv:2012.13666 (2020).

1.6 Organization of the Thesis

The thesis is organized as follows:

- **Chapter 1: Introduction** explains the motivation behind this research, followed by a description of SR with deep learning. Then, the contribution of this thesis is presented, and finally, the publications and submissions during the M.Sc. program are listed.
- **Chapter 2: Review of SR with Deep Learning** provides a review on popular deep learning architectures for SR, including CNN and GAN. The latest advancements on face SR and MRI SR are also reviewed.

- **Chapter 3: Capsule GAN for Robust Face Super-resolution** proposes a robust Multi-scale Gradient Capsule GAN for face SR. Then expand this model to enhance its performance. The performance of this model is evaluated and compared with the state-of-the-art works.

- **Chapter 4: MRI Super-resolution** explained the steps for applying the model proposed in Chapter 3 to a medical application. The architecture of prostate SR model is reviewed. Then, the dataset and the preprocessing pipeline is explained followed by the experimental results and comparison with related works.

- **Chapter 5: Conclusions and Future work** summarizes this thesis and discusses potential future works.

2. Review of Super-resolution with Deep Learning

In this chapter, a review of the Convolutional SR models are presented and the performance of the most powerful CNN models are compared. Then, some of the best GAN-based models for general SR are introduced. Face SR as a sub-domain of SR has been a trendy topic in the last decade. These face SR models are explained in this chapter. Finally, SR in medical imaging is discussed and the previous studies in this field are introduced.

2.1 CNN for Super-resolution

SRCNN is a three-layer convolutional neural network utilized for SR. This shallow and simple network is being trained with MSE loss function [45]. The first model using the normal deconvolution layer is FSRCNN [46]. This model benefits from the deconvolution layer since this layer reduces the computational complexity notably. An efficient sub-pixel convolutional layer was proposed by Shi *et al.* which is called ESPCN [17]. Unlike ordinary deconvolution, in the ESPCN layer, the dimension of the channel is increasing for the purpose of image enlargement. As a result, a smaller kernel size is sufficient. Hence, the computational complexity and training time can be reduced notably.

It has been shown that deeper models with more layers can perform better in many tasks, including SR [47]. VDSR is the first very deep model for SR [18]. This network uses the VGG architecture and has 20 layers. This model is benefiting from multi-scale SR and residual SR as well. Each image is super-resolved for many scales and the network is reconstructing the high-frequency information and adds it to the bicubic interpolation of the LR image. To reduce the number of parameters in VDSR, DRCN was introduced [19]. DRCN utilizes a

recursive convolutional layer 16 times. To overcome the difficulty of training this model, a multi-supervised learning strategy is implemented. In other words, the final result can be considered as the result of the fusion of all 16 intermediate outputs. Since ResNet surpasses VGG architecture in many tasks, it became an interesting choice for SR [48]. SRResNet was proposed as the first ResNet for SR [43]. This model is using 16 residual units followed by a batch normalization to stabilize the training process. Le *et al.* proposed EDSR which is currently the state-of-the-art general SR model [21]. The main difference between EDSR and SRResNet are first, the batch normalization layers were removed, second, the number of output features were increased, and third, the weights for high-scale SR is initiated based on $\times 2$ SR weights.

The performance of the aforementioned convolutional SR models alongside some other famous SR models is compared in terms of PSNR in Fig. 2.1.

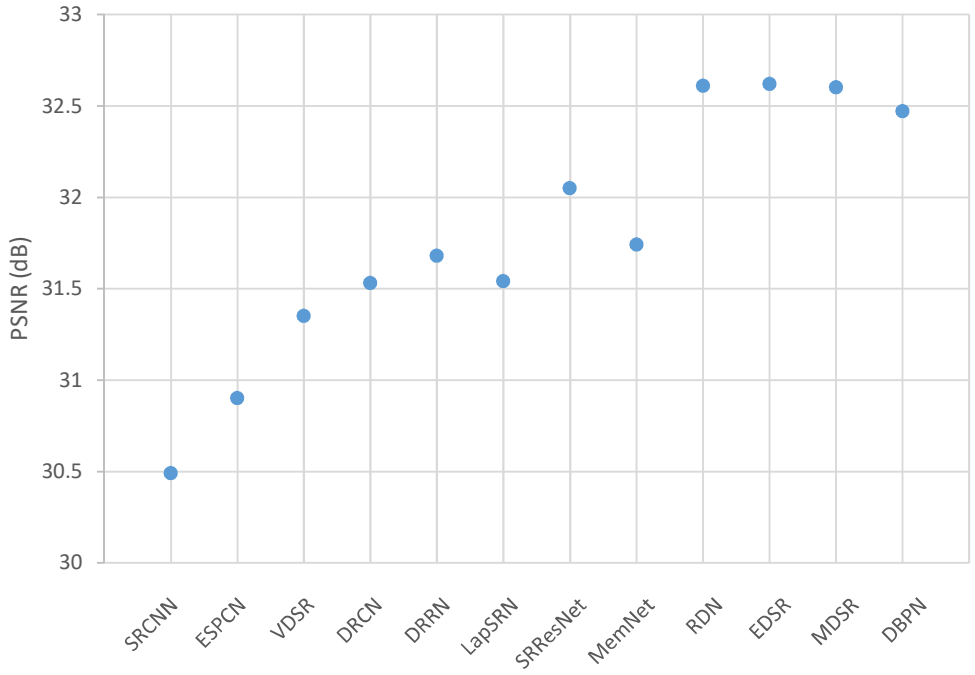


Figure 2.1: The PSNR of different convolutional SR models through time (from 2016 to 2019) [2].

Through time, the models are getting more complex and the PSNR of the outputs are improving. Fig. 2.2 depicts the number of parameters in these models.

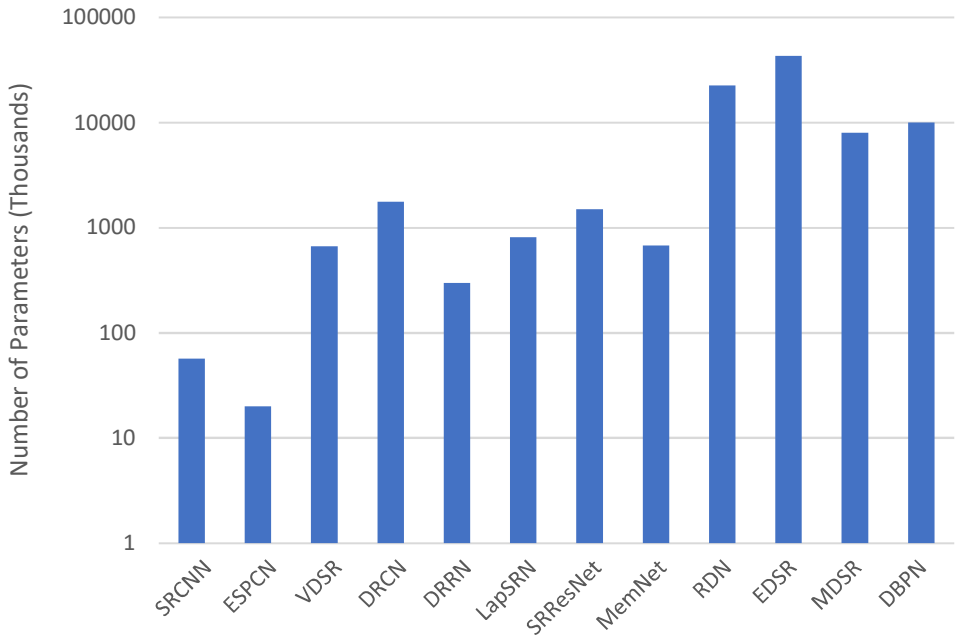


Figure 2.2: Number of parameters of different convolutional SR models through time (from 2016 to 2019) [2].

From 2016 to 2019, the number of parameters increases more than 750 times from 57k to 43m. More powerful hardware for training and more efficient training algorithms make this advancement possible.

2.2 GAN Based Super-resolution Models

GAN-based SR systems are capable of learning to generate super-resolved images with a similar distribution to the real samples. As a result, the outputs are more realistic and visually pleasing in comparison with CNNs. SRGAN is the first GAN-based architecture used for SR [43]. One of the main issues of the convolutional SR model is the dependency of the loss on pixel values rather than the content. In order to tackle this problem, the perceptual loss is utilized in SRGAN training. This loss function evaluates the similarity of the content between the fake generated image and the ground truth. ESRGAN was proposed to improve SRGAN architecture [49]. This model is benefiting from residual-in-residual dense block (RRDB) in the architecture of its generator. The batch normalization layer has been removed as well. A novel loss function, representing the texture similarity between generated

images and real samples is used which improves the output quality.

2.3 Face Super-resolution

Zhou *et al.* proposed the first CNN for face SR [50]. The features are extracted from the LR image and combined through a fully connected layer to form the HR image. This model uses 48×48 images in the output. TDAE, an auto-encoder-based architecture, was proposed by Yu *et al.*. This model is trained to deal with noisy unaligned images [40]. TDAE is basically a decoder-encoder-decoder architecture. The image is up-sampled by the decoder. Then the noise-free up-sampled image is passed to encoder-decoder architecture to form a scale of $\times 8$ super-resolved output. Grm *et al.* proposed CSRIP which is a cascaded progressive CNN model [51]. A face recognition model is utilized alongside the model as well to enhance the performance of CSRIP.

To improve the visual quality of the outputs, the GAN-based models have been used recently in the face SR models. URDGAN, proposed by Yu and Porikli, is a factor of 8 face SR model trained with competitive learning paradigm [38]. Xu *et al.* improved the quality of the outputs by proposing a new texture loss and using labeled data for the training [52].

Various face SR models employ facial attribute domain information to enhance the visual quality of the outputs, especially the facial details and expressions. This information can be obtained from a facial feature extractor embedded in the GAN architecture. Yu *et al.* proposed semantic clues which lead to its supremacy over previous works [53]. Li *et al.* utilized two networks in order to embed the facial attributes extracted from the LR image to the SR model [54]. Another approach for facial attribute information extraction is auto-encoder. Variational Auto-Encoder (VAE), introduced by Liu *et al.*, is used to extract facial attributes from intermediate results [55]. The final output is generated by another convolutional network benefiting from the provided information. Chen *et al.* used landmark heat-map for training their CNN [3]. The architecture of their model is presented in Fig. 2.3.

This facial prior enables their model to focus more on facial details by increasing the

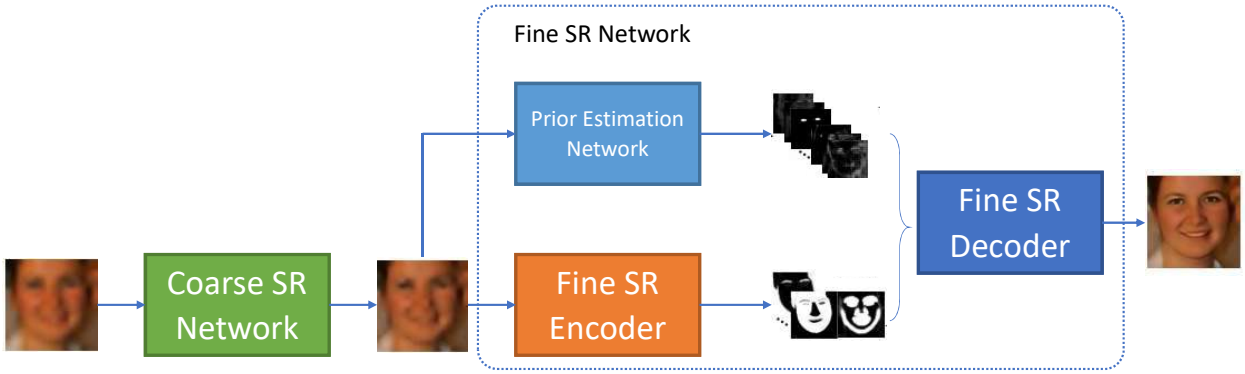


Figure 2.3: The architecture of FSRNet [3].

penalty for error in these areas. Another model that uses two CNNs is proposed by Wang *et al.* [56]. First, ParsingNet obtains prior knowledge from the LR image. Then, FishSRNet reconstructs the HR image benefiting from the provided information. Finally, Ge *et al.*, employed teacher-student concept for face SR [57]. The teacher is trained on the complex domain which is HR and then transfers its knowledge to the less complex model, the student. So the student could imitate the behavior of the teacher on the LR domain.

2.4 MRI Image Super-resolution

SR in medical imaging has been around for a while. The early works used image processing algorithms for medical SR. Rousseau *et al.* employed anatomical inter-modality priors obtained from a particular reference sample [58]. Peeters *et al.* interpolated slice-shifted images in order to enhance signal to noise ratio [59]. Moreover, the effective slice thickness was decreased as well.

More recently, machine learning algorithms have become more popular for medical SR. A 2D convolutional network has been used by Yang *et al.* and Park *et al.* [60, 61]. 3D SR was addressed by Chaudhari *et al.* and Chen *et al.* using 3D architectures [62,63]. These machine learning approaches surpassed traditional image processing algorithms in both visual quality and computational complexity.

Among all machine learning algorithms, GAN performs the best for SR [64]. A GAN using 3D convolutional layers is proposed by Li *et al.* for slice thickness reduction in MRI.

SRGAN, as one of the first GAN models for SR, was utilized by Sood *et al.* for normal and anisotropic SR in prostate MRI [65]. Their findings suggest that the visual quality of the super-resolved images obtained by SRGAN is notably better than other approaches.

3. Capsule GAN for Robust Face Super-resolution [1]

Face hallucination is an emerging sub-field of SR which aims to reconstruct the HR facial image given its LR counterpart. The task becomes more challenging when the LR image is extremely small due to the image distortion in the super-resolved results. A variety of deep learning-based approaches have been introduced to address this issue by using attribute domain information. However, a more complex dataset or even further networks are required for training these models. In order to avoid these complexities and yet preserve the precision in reconstructed output, a robust Multi-Scale Gradient capsule GAN for face SR is proposed. A novel similarity metric called Feature SIMilarity (FSIM) is introduced as well. The proposed network surpassed state-of-the-art face SR systems in all metrics and demonstrates more robust performance while facing image transformations.

3.1 Background

In this section, a brief review of capsule network architecture is presented. Afterward, the concept of GAN and its applications are discussed.

3.1.1 Capsule Network

Capsule Network consists of computational units called capsules. Each capsule is a group of neurons nested together, which represent the substantiation parameters of a particular feature by using a vector. This vector represents the pose parameters of the object. The

1.Most of the parts in this chapter have been published in Majdabadi MM, Ko SB. "Capsule GAN for robust face super-resolution." Multimedia Tools and Applications 79.41 (2020): 31205-31218.

length of the vector corresponds to the probability of the presence of that particular feature. Each vector is multiplied by a weight matrix to predict the vectors corresponding to each higher-level feature. Then, dynamic routing evaluates the agreement of these predictions and computes the final vector for each capsule in the second layer. Fig 3.1 indicates the structure of two typical Capsule layers where m is the number of capsules in the first layer, d_1 is the size of each capsule in the first layer, n is the number of capsules in the second layer, and d_2 is the size of each capsule in the second layer.

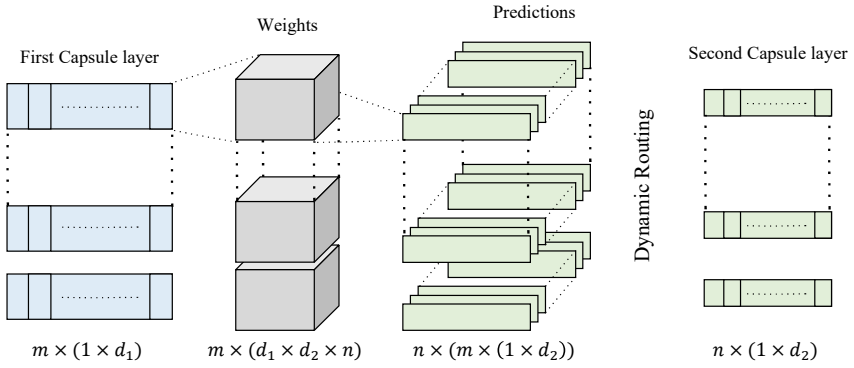


Figure 3.1: Two capsule layers

Unlike CNN, CapsNet is capable of learning the hierarchy of the features and the geometrical relationship between objects. This unique feature makes CapsNet a pose invariant network. As a result, it has surpassed CNN in some classification tasks in terms of accuracy and learning speed [66]. When a CNN is trained with a group of images in which a particular object always appears in a specific pose, the network is incapable of detecting the very same object with a different pose. The reason is that the CNN has only learned to detect the object only based on the presence of the combination of certain features and when the pose of the object has been changed, the low-level features of the image such as edges differ inevitably. Hence, this network is not robust to variations in the pose of the object. To increase the robustness of the SR system, instead of using a bigger dataset or applying data augmentation, CapsNet is utilized as the discriminator in the proposed SR network.

3.1.2 GAN

Similar to the animal learning behavior, GANs can learn through competition [67]. GAN is using two deep learning models called generator and discriminator. The generator is responsible for data generation and the discriminator is classifying samples into real and fake classes. The discriminator is being trained to distinguish between the output of the generator as synthesized samples and the real data and the generator's goal is to fabricate a data sample to deceive the discriminator. Fig. 3.2 depicts the architecture of a basic GAN.

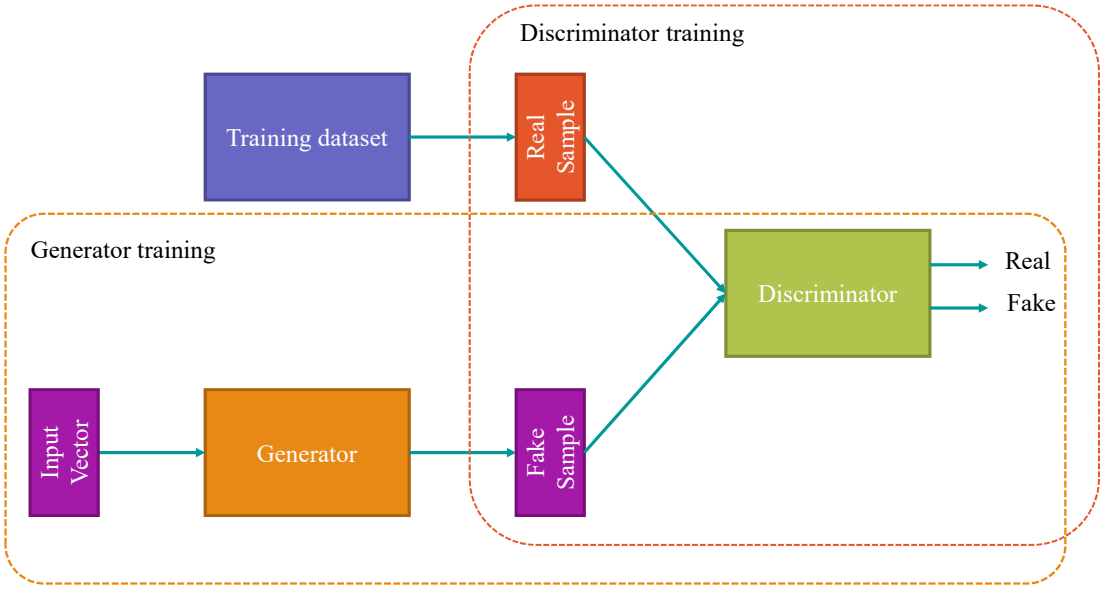


Figure 3.2: The architecture of a typical GAN.

Maximizing the accuracy of the fake and real image classification is the goal of the discriminator while the generator tries to minimize this accuracy by generating better fake samples. It has been demonstrated that GAN is capable of imitating the distribution of the data [31].

One of the challenges in training GANs is mode collapse. Mode collapse is a status of the GAN when the model is not able to generalize. As a result of this error, the generator is only producing the same or very similar outputs. Many solutions have been suggested in order to overcome this issue, such as label smoothing. Label smoothing is a smart way to prevent the discriminator from overfitting. By assigning 0.9 to one class and 0.1 to the other class, the discriminator will face a penalty when it detects a fake or real sample with

too much confidence (1 or 0). As a result, it can establish a more general understanding of the real and fake samples. Noisy labels can improve the label smoothing process. A random number in the neighborhood of 0.9 and 0.1 are assigned to real and fake classes, respectively. As a result, the discriminator can't overfit on the exact labels.

3.2 MSG-CapsGAN

GAN has surpassed other deep learning approaches in many image-related applications, especially SR task [53]. GAN mainly consists of two networks that are trained competitively, Generator, and Discriminator. The high-level demonstration of the proposed network is depicted in Fig. 3.3.

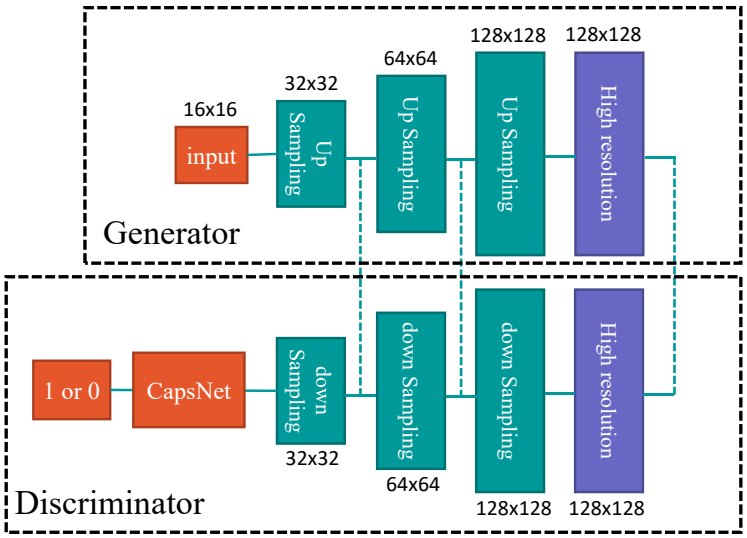


Figure 3.3: High level illustration of MSG-CapsGAN.

This architecture is called Multi-scale Gradient GAN (MSG-GAN), which is used for the SR task for the first time in MSG-CapsGAN [44]. These intermediate connections can be used as an alternative for progressive training, so great results have been acquired from this architecture for synthesizing high-resolution images [68]. These promising results were the primary motivation for using this idea for $\times 8$ SR. The supremacy of this architecture over other networks has been illustrated in our previous work [44].

As far as the discriminator is concerned, in our work, CapsNet is utilized. CapsNet benefits from some advantages over classic classifiers, especially, being pose-invariant [66].

The input of the discriminator is either the super-resolved image or HR image, and its output is 1 or 0 representing real or fake. First, the input image is down-sampled by using three convolutional layers with the stride of 2 and the activation function of the leaky Rectified Linear Unit (ReLU). Then the output is applied to a CapsNet with two layers. The primary capsule layer has 32 capsules with 8 neurons in each capsule. The last layer contains 10 capsules with 16 neurons. Dynamic routing is performed for three iterations. Finally, the output of CapsNet is connected to the output neuron with a fully connected layer. The activation function of the last layer is the sigmoid function, with the output in the range of 0 to 1.

The second network is the generator. The 16×16 pixel input is applied to the generator. The image is up-sampled 3 times in order to reach 128×128 size. The architecture of the up-sampling unit is depicted in Fig. 3.4.

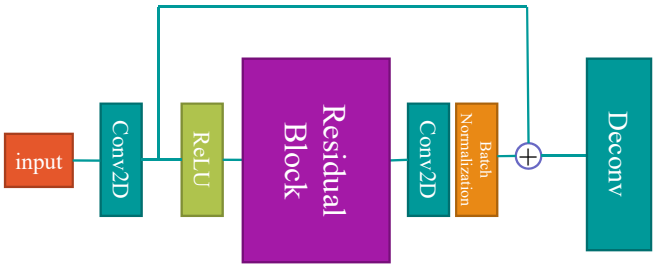


Figure 3.4: The architecture of up-sampling unit.

The first convolutional layer has 64 filters with a kernel size of 3×3. Then, ReLU is applied. After the residual unit, a convolutional layer with the same parameters as the first Convolutional layer is presented, and it is followed by a batch normalization layer. The output is integrated with the output of the first convolutional layer. Finally, a deconvolution is performed on the output results. The layer structure of the Residual Block is exhibited in Fig 3.5.

All convolutional layers have 64 filters and the same padding. The input is added to the output of the last layer to form the output of the residual block.

By using 2RGB layers after intermediate connections, a multi-scale SR model can be acquired. These layers are basically 3-filter convolutional layers, where each filter represents

Figure 3.5: The architecture of the residual block.

one channel in the output tensor. So rather than only using these connections for passing the information between the generator and the discriminator layers, they can also be utilized for multi-scale training. Eq. 3.1 shows the perceptual loss function used for training the generator.

where l^{SR} is the training loss, S is a group of different scales which is $\{32, 64, 128\}$, l_i^{SR} is the content loss for each scale, and l_{GAN}^{SR} is the GAN loss. The content loss is computed as follows:

$$l_i^{SR} = \frac{(f_{VGG19}(x_i, 9) - f_{VGG19}(y_i, 9))^2}{N} \tag{3.1}$$

where $f_{VGG19}(x_i, 9)$ and $f_{VGG19}(y_i, 9)$ are the outputs of the ninth layer of VGG19 network with the input of the generated image and the true image with the scale of i, respectively and N is the number of elements in this feature vector. Both the HR image and super-resolved image are passed to the VGG19 network and the similarity of the vector feature in the $9th$ layer of the network is evaluated using Mean Square Error (MSE).

In the next step of improving MSG-CapsGAN, the network has been forced to learn easier steps first, inspired by the progressive training. Then, the task is gradually getting more complex. In order to implement this training approach, the loss function has been modified as follows:

$$l^{SR} = \sum_{i \in S} a_i \times l_i^{SR} + 3E(-3) l_{Gen}^{SR} \tag{3.2}$$

The hyperparameter a_i shows the state of the network in the training and controlling the contribution of each error to the final loss. These hyperparameters are adjusted according to the Fig 3.6.

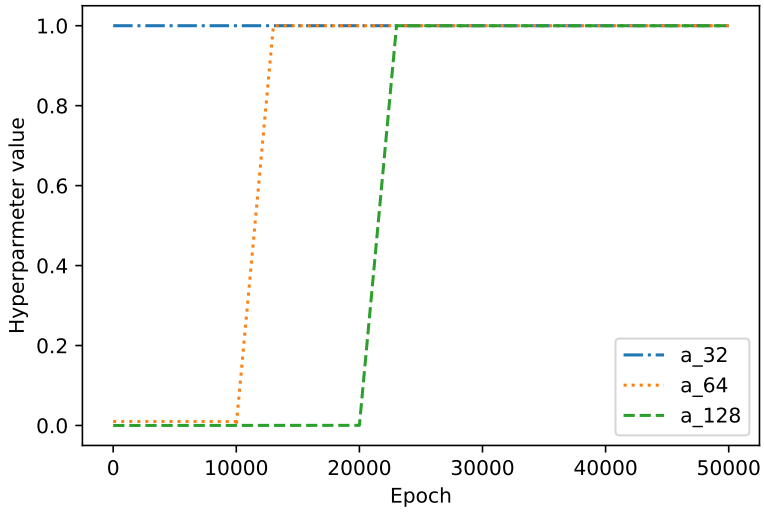


Figure 3.6: The value of hyperparameters through training.

At the beginning, a_{32}, is equal to 1, and all other hyperparameters are equal to 0. Hence, the only content loss that matters is for the scale of 32. Then, the contribution of the a_{64} and a_{128} increases gradually.

Studies demonstrate that using a matrix as a representation of being real or fake can improve the performance of the GAN notably. This network is called patch GAN [69]. Inspired by this idea, the output of the discriminator is redesigned. A fully connected layer with 4900 neurons is connected to the last capsule layer. Each value of each capsule is connected to all of the neurons through a weight. Then, the 4900 neurons are rearranged to 70×70 array of values in the range of 0 to 1, where 0 represents fake, and 1 represents real.

In order to further reduce the artifacts in the output image, two approaches are taken. First, fully residual architecture is utilized. These types of networks are proven to be effective for image correction and SR [70]. Second, transfer learning can boost the performance of the proposed network since very informative features can be extracted using pre-trained models. Fig. 3.7 shows the diagram of the proposed architecture.

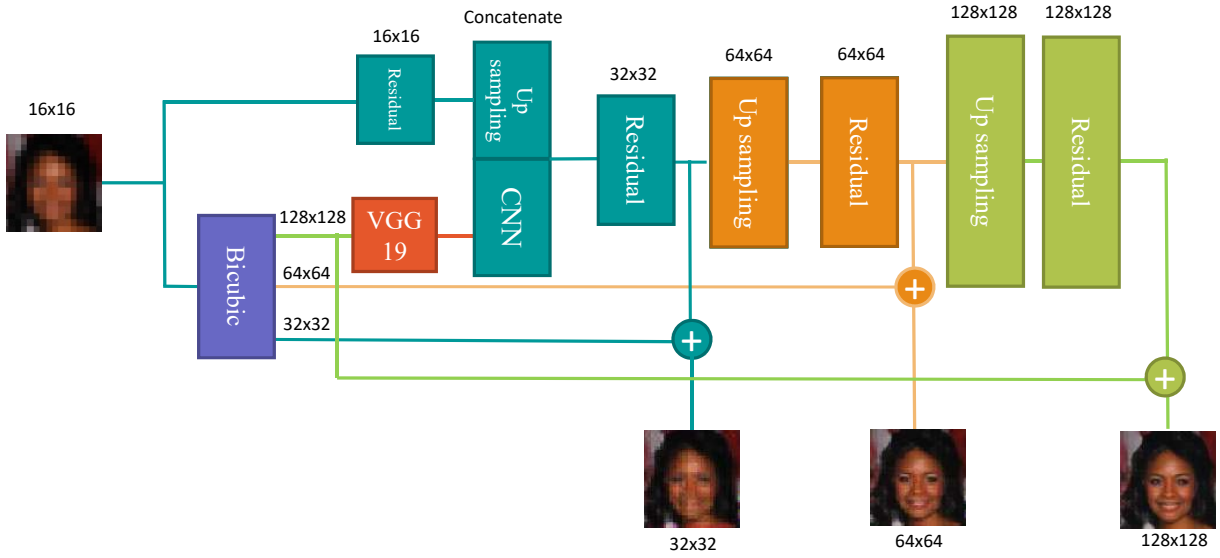


Figure 3.7: High level illustration of the proposed VGG_Residual network.

The input image is up-sampled using the bicubic method, and then the network is reconstructing the high-frequency details of the image. This high-frequency information is added to the up-sampled image to form the output. It is also passed to VGG19 layers, and the feature map is extracted. Then, the extracted feature map is concatenated with the convolution layers of the network so that the network could benefit from these features as well.

3.3 Experimental Results and Discussions

The model is implemented using Tensor flow version 1 alongside with KERAS package. The language used in this thesis is python. The codes for MSG-CapsGAN is publicly available on GitHub [1]. The new layers implemented for this project are also published in the same GitHub repository.

3.3.1 Dataset and Preprocessing

To investigate the capability of the proposed network for facial SR task, aligned CelebA dataset [71] is used in this study. CelebFaces Attributes Dataset (CelebA) is a large dataset

[1] https://github.com/MahdiyarMM/MSG-CapsGAN

of face images. The dataset is annotated with 40 binary labels such as mustache, eyeglasses, smiling, etc. There are also 5 landmark locations provided. The images are collected form the internet. Recently, the CelebA mask dataset is released as well. The dataset can be accessed online [2]. This aligned dataset has been used so the robustness of the network could be evaluated for different poses and angles of the faces later. The dataset is divided into training and test-set with 162,770 and 19,867 images, respectively. Fig 3.8 exhibits some samples from this dataset.

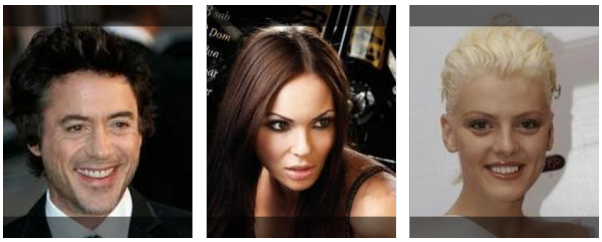


Figure 3.8: 3 samples from CelebA aligned dataset.

The size of each image is originally 218×178. As it is shown in Fig. 3.8, 20 pixels from the top and 20 pixels from the bottom of each image are cropped. Hence, the final images are squares with a size of 178×178. Cropping the images reduces the background and enables us to compare our results with similar works [4]. Finally, all the images are resized to 128×128 as HR image and 16×16 as LR images. No further enhancements or adjustments are applied to the dataset samples.

3.3.2 Performance Metrics

In order to evaluate the performance of the SR networks, three metrics are mostly used:

1. PSNR: Peak Signal to Noise Ratio

$$PSNR = 10 \log_{10} (\frac{L^2}{MSE}) \tag{3.3}$$

where L is the maximum value in the image and MSE is the Mean Square Error of the super-resolved image.

[2] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

2. SSIM: Structural SIMilarity [72]

$$SSIM(I, \hat{I}) = \frac{2\mu_I\mu_{\hat{I}} + k_1}{\mu_I^2 + \mu_{\hat{I}}^2 + k_1} . \frac{\sigma_{I\hat{I}} + k_2}{\sigma_I^2 + \sigma_{\hat{I}}^2 + k_2} \tag{3.4}$$

where μ is the mean, σ is the variance, I and $\hat{I}$ are input images, $\sigma_{I\hat{I}}$ is the covarience between I and $\hat{I}$, and k_1 and k_2 are two constants.

3. MS-SSIM: Multi-Scale Structural SIMilarity [73]

PSNR is a pixel-wise metric representing the similarity of the corresponding pixel values in two images. However, SSIM and MS-SSIM measure the quality of the reconstruction of the structural information in the output image. These two metrics heavily rely on the statistical properties of the images so they can estimate the similarity of two images from the human perception point of view more accurately. Note that it has been demonstrated that MS-SSIM can approximate human perception more accurately than SSIM and PSNR [73].

The output of an SR system can be potentially used as an input of another deep learning-based system such as semantic segmentation, object detection, and face recognition. Hence, it is important to investigate the similarity in the informative features of the super-resolved image and the ground truth. To have a closer metric to current deep learning understanding for image similarity, a new simple but powerful metric is proposed, which is called Feature SIMilarity (FSIM). There are some particular features in each image that carry crucial information. These features are combinations of more basic visual characteristics such as edges and brightness. It is critical to pay more attention to these informative features. One way to obtain this data from the image is transfer learning. A pre-trained network that is designed to classify millions of images into thousands of classes can extract these features. By computing the similarity of the extracted features from the original image and the reconstructed image, the similarity of the two images can be evaluated. Eq. 3.5 explains how FSIM is computed.

$$FSIM = e^{-C \times NRMSE(f_1, f_2)} \tag{3.5}$$

where f_1 is the feature matrix of the reconstructed image, f_2 is the feature matrix of the source image, C is a constant to adjust the distribution of the index properly. $NRMSE$ is the normalized root mean square error computed as follows:

$$NRMSE(f_1, f_2) = \frac{\sqrt{\frac{(f_1 - f_2)^2}{N}}}{\bar{f}_2} \tag{3.6}$$

where $\bar{f}_2$ is the mean of f_2 and N is the number of values in feature matrices.

FSIM is always in the range of 0 to 1, which makes it easy to understand. By increasing the similarity of the reconstructed image to the source image, the FSIM gets closer to 1. Despite the variation of size and architecture of pre-trained available networks, almost any network can be used for extracting f_1 and f_2 and only C should be adjusted according to the output values in the feature vector for proper distribution of the matrix. In this thesis, VGG16 [74] is used for computing FSIM and C is set to 0.3.

3.3.3 Results

As explained earlier in section 1.4, different modification has been applied to the network proposed in [44]. To investigate the contribution of each modification on the performance of the network, PSNR, and SSIM of the network in each step are evaluated. Table 3.1 presents the parameters corresponding to these model evaluations.

Table 3.1: Experimental parameters and details.

Experimental Parameters						
Training			**Batch size**		**Label smoothing**	
Epochs	**Time**	**Hardware**	**Full**	**Half**	**Positive**	**Negative**
50K	52.8 h	Tesla K40 (12 GB)	32	16	[0.7-1]	[0-0.3]

As addressed in Table 3.1, label smoothing is applied to prevent overfitting. In each training iteration, a random label is generated for each class with normal distribution and the aforementioned range. Moreover, the full batch is used for training the generator and the

discriminator is trained utilizing half batch. It is worth mentioning that the optimizer for training discriminator is Gradient descent (with momentum) optimizer (SGD) with the learning rate of 0.1 and decay rate of 10^{-6}. Regrading the generator, the model is trained with ADAM with the learning rate of 2×10^{-4}.

The performance of the models in each step of the modification is evaluated and the results are summarised in Table 3.2.

Table 3.2: The results

Method	**PSNR**	**SSIM**
MSG-CapsGAN [44]	23.35	0.673
Proposed Multi-scale SR	23.50	0.6691
Proposed Progressive a_i adjustment	23.54	0.673
Proposed Patch GAN	***23.64***	0.717
Proposed VGG_Residual	23.53	***0.719***

MSG-CapsGAN is the model that we have introduced in [44], Multi-scale SR is MSG-CapsGAN with 2RGB layers, Progressive a_i adjustment is using Eq. 3.2 as the loss function, Patch GAN has a 70×70 output, and the architecture of VGG_Residual network is depicted in Fig. 3.7. VGG_Residual network achieves the highest SSIM. However, the PSNR has been reduced for 0.11 in comparison with Patch GAN. A more detailed comparison between the two best networks will be presented in the next subsection.

3.3.4 Comparison with State-of-the-Art

The similarity of the super-resolved images with ground truth in different models is evaluated and summarized in Table 3.3.

The proposed models surpassed Progressive Face SR [4] in terms of SSIM, MS-SSIM and FSIM and outperforms VDSR [18] in terms of PSNR. Hence, the proposed VGG_Residual CapsGAN outrun state-of-the-art models in all four matrices with a notable margin, especially in MS-SSIM and FSIM, which are better representatives for human and AI understanding of the similarity of two images, respectively. Another precise face SR model is FSRNet [3]. The findings of [4] demonstrate the primacy of Progressive Face SR over FSR-

Table 3.3: Comparison of the performance of different SR systems.

Method	**PSNR**	**SSIM**	**MS-SSIM**	**FSIM**
Bilinear	20.75	0.574	0.782	0.5320
Progressive Face SR [4]	22.67	0.687	0.908	0.6374
VDSR [18]	22.96	0.655	0.887	0.6103
MSG-CapsGAN [44]	23.35	0.673	0.899	0.6371
Proposed Patch GAN	***23.64***	0.717	0.927	0.6788
Proposed VGG_Residual	23.53	***0.719***	***0.929***	***0.6918***

Net. Hence the proposed VGG_Residual model surpassed this network as well with a margin of 0.89 for PSNR, 0.078 for SSIM, and 0.082 for MS-SIM.

As explained earlier, many studies have used attribute domain information for dealing with image distortion in face SR by using Facial Attribute Networks (FAN) and heatmap loss. Some researchers have utilized growing networks as well. These approaches demonstrate promising results although they increase the complexity of the network notably. Moreover, their application will be limited to face-related tasks, because of employing FANs. Table 3.4 demonstrates a comparison between the proposed model and state-of-the-art face SR systems.

Table 3.4: Comparison of the complexity of different face SR systems.

Model	**Growing Architecture**	**Heatmap Loss**	**FAN**
FSRNet	No	Yes	Yes
Progressive Face SR	Yes	Yes	Yes
Proposed model	No	No	No

The superiority of the proposed model is acquired with a less complex model and training process. Unlike other face SR models, the proposed architecture is not using any attribute domain information or growing architecture. As a result, the application of the proposed model is not limited to face-related tasks.

A visual comparison between the two best-proposed models and the state-of-the-art face

SR system is depicted in Fig. 3.9 for four test samples.

(a) (b) (c) (d) (e) (f)

Figure 3.9: (a)16 × 16 input (b)Bilinear (c)Progressive [4], (d)Proposed Patch GAN, (e)Proposed VGG_Residual, and (f)High resolution 128 × 128.

It can be seen that the outputs of [4] are sharper. However, the proposed SR methods generate more similar images to the high-resolution target images. Moreover, the VGG_Residual network reconstructs the facial details such as eyes more clearly in comparison with Patch GAN. Between Patch GAN and VGG_Residual network, the former achieved better PSNR, but the latter outperforms Patch GAN in all other metrics. The computational complexity of VGG_Residual is more than Patch GAN because of the embedded non-trainable VGG19 and bicubic layer. So, there is a trade-off between complexity and performance between these two models.

3.3.5 Robustness Test

As was mentioned in section 3.1.1, one of the advantages of using CapsNet over CNN is its robustness. CapsNet is capable of learning the hierarchy of the features, which makes it a pose invariant network. Robustness is a quality of system to deal with the unprecedented

scenarios. One of the most common meanings of robustness is that the system can exhibit good performance when facing a noisy input. However, the robustness has broader meaning. In this thesis, image transformation is used to evaluate the robustness of the model. The network is trained on aligned dataset. All the samples are aligned in the way that the eyes are at the center of the image and all people are facing camera. Hence, the rotated and scaled image is an example of unprecedented scenario. The robustness of the proposed network is compared with [4]. The test set samples are rotated and scaled simultaneously. Then, the transformed LR images are applied to both SR systems. Fig. 3.10 depicts an illustration of the mentioned transformation.

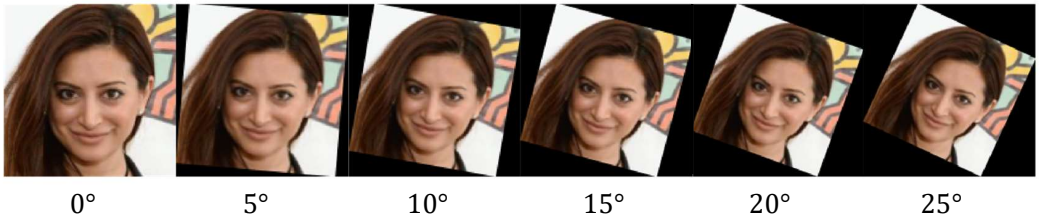


Figure 3.10: The transformation used in the robustness test.

It is worth mentioning that images are flipped randomly, so the rotation is from $-25°$ to $25°$. In the last step, the percentage of the drop in each metric is calculated as it is displayed in Fig. 3.11.

The proposed network outperforms [4] in all metrics with a notable margin except in SSIM. Regrading the robustness in SSIM, for small variation in the images, similar behavior can be witnessed from both networks. For greater angles, [4] have better performance. However, the SSIM index value is still higher in the proposed design. Furthermore, as was explained in section 3.3.2, MS-SSIM can provide a better approximation for human perception [73]. Moreover, for deep learning-based applications, FSIM can express the similarity more precisely. In both of these metrics, the proposed design surpassed [4]. As we expected, these results imply that using CapsNet as the discriminator of the SR GAN system can enhance the robustness of the network.

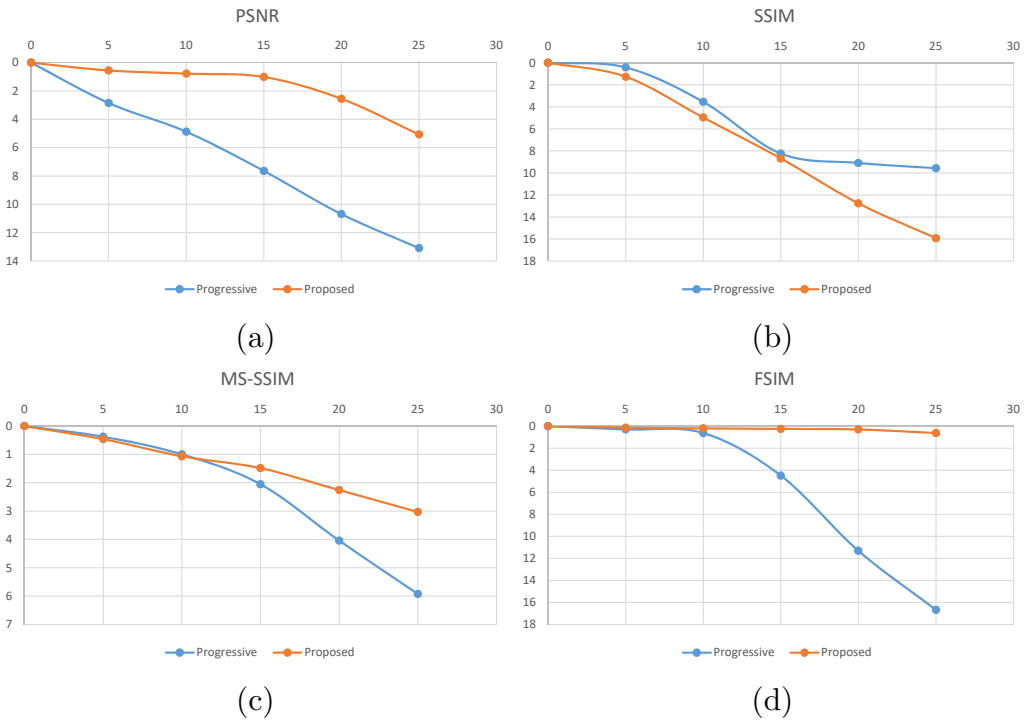


Figure 3.11: The percentage of drop in (a) PSNR, (b) SSIM, (c) MS-SSIM, and (d) FSIM vs the rotation angle.

3.4 Summary

There are several face-related tasks that can benefit from higher resolution images. For example, in surveillance systems, because of the number of cameras, the quality of each camera is not very high. Hence, many footages obtained from these systems do not have adequate quality for some important tasks such as identification. When the resolution of a face is too low, it is impossible to identify the person, since the facial details are not visible. In this case, AI can increase the resolution of image and reconstruct the facial details required for identification. As a result, without changing the already established infrastructure, an impossible task can become possible. This system can improve the performance of face-related systems without any cost imposition in the hardware and infrastructure.

To develop an accurate, efficient, and powerful SR system that could address a variety of applications, a robust and general SR system is proposed. The main goal in this thesis indeed is to propose a face SR system but a general SR system that could work on other types of data is more beneficial and useful. However, this means no attribute domain information

or labeled data can be utilized in this thesis. Although many previous works in the field of face SR, employed facial attribute data or facial detail extractors in the training process and achieve very good results, we are addressing the quality problem with a different approach in order to have a powerful yet general model. Instead of providing more data for the proposed model to help the model learn more efficiently, more powerful architecture is used. Hence, the model is trained unsupervised with unlabeled data.

GAN is known for its unique ability to fabricate samples with the same distribution of samples in a real dataset. Relying on this unprecedented ability, GAN-based models can generate hyper-realistic samples. So they soon became the most popular architecture in SR systems. However, traditional GANs suffer from artifacts, image distortion, and mode collapse, especially when attempting to generate very high resolution outputs. As a result, in high-scale SR (such as $8\times$) traditional GANs are not able to perform with a satisfactory result. In order to address this issue, many ideas have been proposed such as progressive growing GAN where the generator and discriminator are growing gradually. Although great results have been achieved with these models, complexity is a concern in growing architectures. Multi-scale Gradient GAN (MSG-GAN) is an efficient alternative solution for high-resolution image generation. This architecture is using intermediate connections to pass gradient information of all scales between the discriminator and the generator. In this thesis, MSG-GAN is employed for SR for the first time. Another significant architecture-wise novelty of this model is using CapsNet in the discriminator. Unlike CNNs, CapsNet learns the relationship between the features. Hence, it can be considered as a pose invariant network. Using this network in the discriminator enhances the robustness of the model. In this thesis, CapsNet is used for SR for the first time. Using MSG-GAN and CapsNet simultaneously causes instability in the training process since both architectures have some level of instability. Choosing the right learning rate and optimizer for the training and balancing the learning process of the generator and the discriminator by using label smoothing and half batch improves the stability, notably.

To even further improve the performance of the proposed MSG-CapsGAN, several modifications have been applied to the model and the performance of each model is compared

so the contribution of each step could be analyzed separately. First, multi-sale SR is implemented. Each intermediate connection in MSG-CapsGAN corresponds to a different scale of SR. In this step, each intermediate result ($2\times$ and $4\times$), as well as final results, are used for calculating the error. Thus, not only the final output but also other scales contribute to the final loss. By implementing this idea, the PSNR and SSIM are improved by 0.15 dB and 0.0039, respectively. In this step, all errors contribute to the loss equally.

In order to make the model learn low-scale SR first, a progressive loss weight adjustment approach is utilized. First, the only error is for $2\times$ SR and after learning this simple task, the weight corresponding to the $4\times$ increases gradually. This process continues until all scales are learned. This process slightly increases both similarity metrics. Furthermore, more stability and fewer artifacts in the final outputs are witnessed in this scenario.

In all of the aforementioned models, the discriminator classifies the whole image as real or fake. More precisely, it estimates the level of being photo-realistic by a value from 0 to 1. Patch GAN outperforms traditional models in generating high-quality visually pleasing images. Hence, the discriminator of MSG-CapsGAN has been modified accordingly. The new output of the model is a matrix instead of a single value. Each member of the matrix, represents the quality of a patch of the image, as small as 1.8×1.8 pixels, in this thesis. This modification improves the PSNR by 0.1 and SSIM with a notable value of 0.044. Using patch GAN might not enhance the pixel-wise similarity much, but it increases the statistical similarity of images notably since it helps the model to concentrate on the problematic areas more.

The final modification in the model is first, using VGG19 for feature extraction, and second, applying the residual learning approach. The first layers of VGG19 are used embedded in the model for feature extraction. The extracted features are used alongside the features extracted by trainable convolutional layers. These features represent the texture and edge so using this model improves the visual quality of the outputs. By training model with residual learning paradigm, the low-frequency information of the image is provided by bicubic interpolation so the generator can only focus on reconstructing high-frequency information of the input image. These two modifications improve the SSIM slightly, although the PSNR drops.

However, as illustrated in Fig. 3.9, the output of this model is more realistic with sharper edges and less texture error. This is another example of the fact that PSNR and SSIM are not able to reflect the true visual similarity of images.

In order to overcome the mentioned issue of similarity metrics, a novel similarity assessment approach is proposed, called FSIM. Inspired by perceptual loss, this metric evaluates the similarity between the features extracted from both images using a proper pre-trained model. In this thesis, VGG 16 is used for this purpose. Another significance of FSIM is that unlike other similarity metrics, its main goal is to evaluate the similarity from the DL perspective, not human perception. In the aforementioned example, the comparison between the proposed patch GAN and the proposed VGG-residual model, the quality improvement is not reflected in PSNR and SSIM. However, a notable increase in FSIM is witnessed. Using all of these metrics, the two best-proposed architectures are compared with state-of-the-art face SR models. The proposed model outperforms other SR systems in all metrics. The margin in FSIM is more noticeable. Another important point is higher PSNR does not necessarily mean better outputs. VDSR [18] has higher PSNR than Progressive Face SR [4], though its outputs have lower quality, as reflected in other metrics, especially FSIM. Visually speaking, Fig. 3.9 illustrates the output of the different models. It can be seen that progressive Face SR [4] can generate very sharp edges and accurate texture in some areas of the image. However, artifact and image distortion are visible in many images. The specific facial features such as eyes have some problems as well. The proposed model performs better in reconstructing symmetric natural facial details.

Since one of the main motivations of this work is developing a general SR model, robustness is an extremely important quality of the model. Since a robust model can be used in various domains and it will perform good in unprecedented scenarios. To evaluate and compare the robustness of the proposed model with state-of-the-art design, the model is trained on the alighted CelebA dataset, and robustness test is conducted. The input image is rotated and scaled and model performance is evaluated. In three out of four metrics, the proposed model performs better, especially in MS-SSIM and FSIM which are a better representation of the true quality of the image.

The proposed model outperforms the state-of-the-art face SR systems in all similarity metrics and shows more robust behavior dealing with image transformation, while no attribute domain information is used and the training process is completely unsupervised. As a result, it can be considered a powerful general SR system. These promising results motivate us to use this model for addressing more complex SR tasks such as medical SR. In the next Chapter, this architecture is employed for prostate MRI SR. Another possibility for expanding this research in the future is to improve the sharpness and the texture of the output image using a more complex loss function. The modified loss function can represent the texture similarity as well as the pixel-wise similarity. Moreover, an edge detection system can be used to increase the penalty in these areas. As a result, the model can concentrate more on the edges and the overall quality of the output can be increased.

4. MRI Super-resolution

Prostate cancer is one of the most common cancers worldwide. The early detection of cancer has a great impact on the survival rate and the success of treatment. One way to increase the chance of early diagnosis is by improving the quality of imaging methods. SR is an effective way to increase the resolution of a scan and reduce the noise level. Motivated by these facts, we have implemented a deep learning model for prostate MRI SR in order to facilitate early diagnosis and help save lives. Currently, there are not many works on high-scale SR in medical scans. The main concern is whether the medical details essential for the diagnosis will be preserved in the process. Recently, powerful DL models such as GANs have improved the performance of the SR models significantly and made high-scale medical SR possible.

4.1 Background

In this section, protest cancer and its diagnosis methods are reviewed. Then, the concept of MSG-GAN and its advantages over similar approaches for HR image generation are explained.

4.1.1 Prostate Cancer

In 2020, the fourth most common cancer in Canada was prostate cancer, as illustrated in Fig. 4.1.

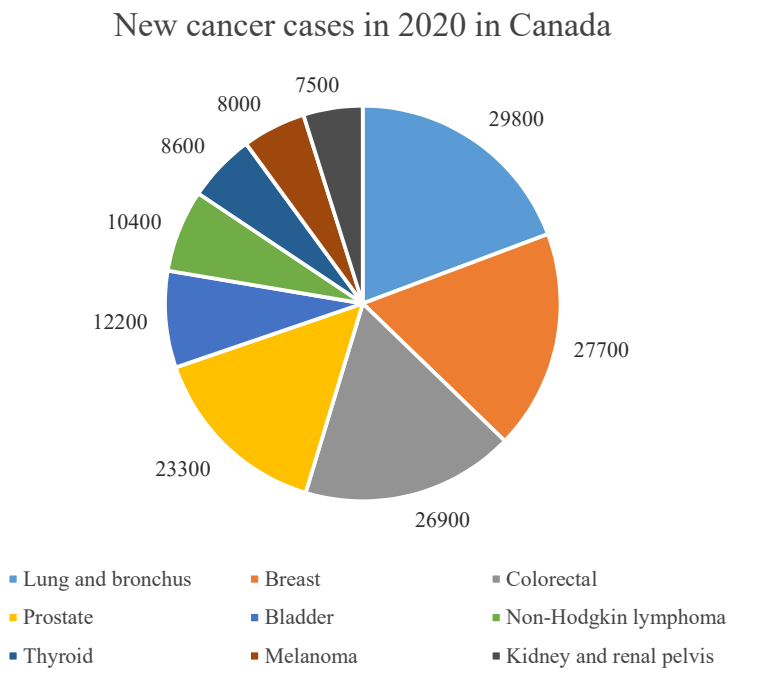


Figure 4.1: The distribution of the new cancer cases in Canada in 2020 [5].

This cancer is identified as the most prevalent non-cutaneous cancer in men. It is also the second-cancer-related cause of death in male individuals. Worldwide, 1 in every 9 men will be diagnosed with prostate cancer through their lifetime [75].

There are two major approaches for monitoring the symptom-free patient for possible prostate cancer:

- **Digital rectal exam (DRE)** is a test in which the doctor examines the rectum for identifying any abnormality in the shape, size, or texture of the prostate.
- **Prostate-specific antigen (PSA) test** is a blood test used for prostate cancer detection. In this test, the density of PSA in the patient's blood is analyzed. A high level of PSA can be an indication of infection, inflammation, or cancer in the prostate.

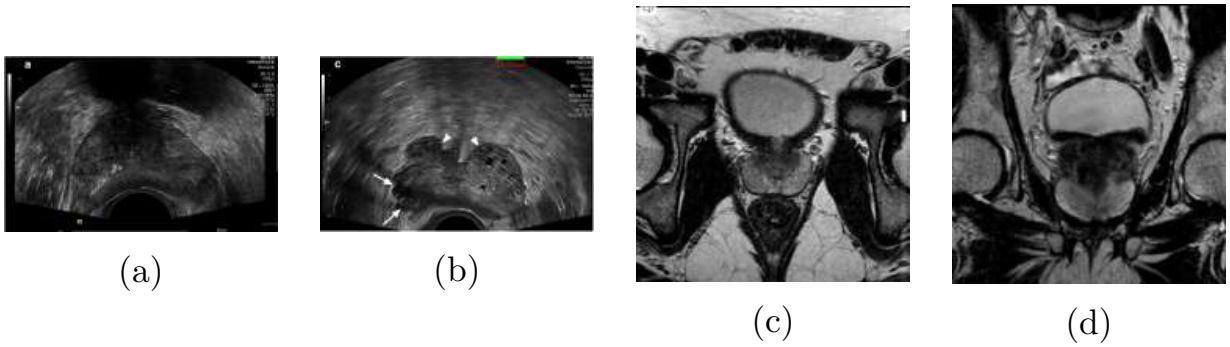
If any abnormality is detected in the aforementioned tests, the following approaches are used for cancer detection:

- **Ultrasound:** A prob is used to take an image of the prostate using ultrasound waves.

- **Magnetic resonance imaging (MRI):** More detailed images of the prostate can be acquired using MRI. This image can be used by the doctor not only for cancer detection but also for treatment planning.

- **Collecting a sample of prostate tissue:** A tissue sample from the prostate is collected through a prostate biopsy. The presence of the cancer cells is investigated in the lab for disease confirmation.

Fig. 4.2 depicts two different types of prostate diagnosis imaging methods.



(a) (b) (c) (d)

Figure 4.2: Samples of ultrasound images of (a) healthy prostate [6], (b) prostate with cancer [6], and the MRI of (a) healthy prostate [7], (b) prostate with cancer [8].

MRI can provide very high-quality images of the prostate. As a result, it is a very helpful way to detect the disease. However, the quality of scans is not always adequate for early-stage cancer detection, especially in cheaper scanning devices. The commonness of prostate cancer and the importance of early diagnosis, alongside the ability of MRI in acquiring good images, motivates us to address MRI SR in prostate cancer.

4.1.2 MSG-GAN

One major challenge in generating a high-resolution image using GAN is the gradient problem. This problem occurs when the distribution of the generated images have not enough overlap with the distribution of the training set. To overcome this issue, a layer-wise solution is proposed [76]. Training GANs progressively is an effective training approach to generate HR images. Fig. 4.3 displays the architecture of progressive GAN.

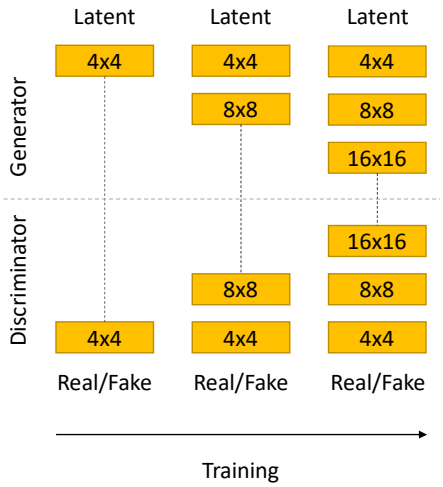


Figure 4.3: The architecture of progressive GAN.

First, both generator and discriminator have a very simple architecture and the model is trained to create LR fake images. In the next step, a new layer is gradually added to both models which enable the GAN to double the output resolution. This process continues to the point that the output size reaches the desired dimension. since the first layers are trained at first, the gradient problem is resolved. However, this architecture and training paradigm is so complex. A simpler yet effective solution called MSG-GAN is proposed by Karnewar *et al.* [77]. This alternative solution attempts to solve the gradient problem using a fixed model, i.e., non-progressive, with new paths for the gradient. Fig. 4.4 depicts the architecture of MSG-GAN.

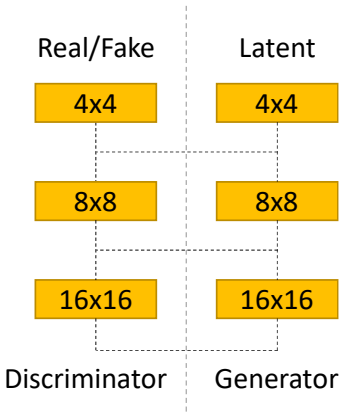


Figure 4.4: The architecture of MSG-GAN.

Unlike progressive GAN, the architecture of MSG-GAN does not change through time. The intermediate connections create new paths for the gradient. In the progressive model, to make sure that the distribution of the fake and real samples in low-scale SR is the same, only the first layer of the models was present in the training. However, in MSG-GAN, the connection between the first layer of the generator and the first layer of the discriminator helps the model to overcome the gradient problem.

In medical SR, reconstructing all medical details useful for diagnosis is crucial. As a result, overcoming the gradient problem in GAN-based SR models is necessary. Motivated by the performance of MSG-GAN, and the flexibility and stability of this architecture, this model is used in this research to form MSG-CapsGAN for the SR problem. More details on the architecture of the proposed GAN for MRI SR are provided in the next section.

4.2 Model Architecture

The architecture used in this study was proposed in our previous work, originally for face SR [24]. This model is the first Multi-scale Gradient Capsule GAN and outperforms state-of-the-art face SR systems in all similarity metrics, as explained in Chapter 3. However, since the dimension of the input image and the type of the data is different, some alterations have been made to the network.

Unlike the face SR problem, the size of the HR image in this study is 224×224. Moreover, each MRI slice is a black and white image with one channel. Another significant difference between this architecture and the architecture in Chapter 3 is the feature extractor.

In MSG-CapsGAN, VGG19 is used in the heart of the model. This model is responsible to extract useful features from the interpolated image and SR system is reconstructing the image using these informative extracted features. However, MRI images are completely different from the samples in the ImageNet dataset. The type of features extracted by VGG can not represent the important content of an MRI scan. Hence, embedding this model in the generator is not useful. Inspired by our previous work with radiography, in this model, VGG is substituted with CheXNet [78].

CheXNet is a deep learning architecture trained for lung disease classification. The architecture of CheXNet is based on DensNet as presented in Fig. 4.5.

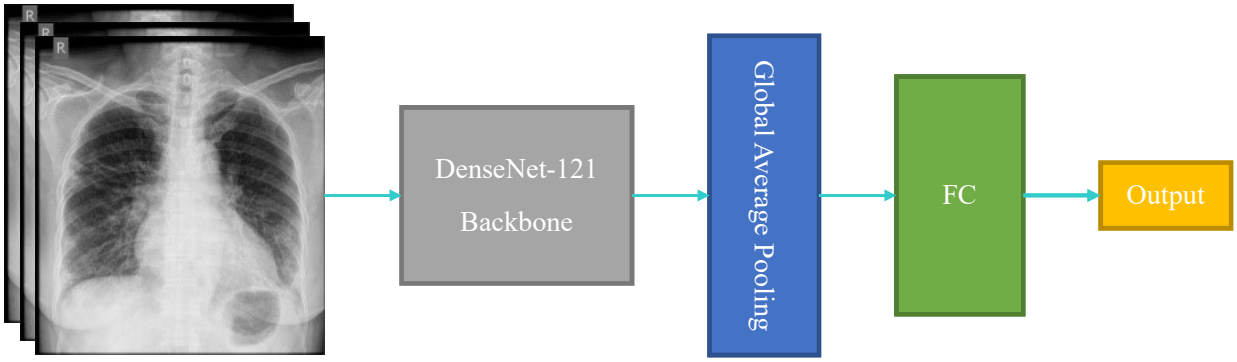


Figure 4.5: The architecture of CheXNet [9].

The input is passed to DenseNet-121, followed by a global average pooling layer. Then, a Fully Connected (FC) generates the output. The architecture of DensNet is shown in Fig. 4.6.

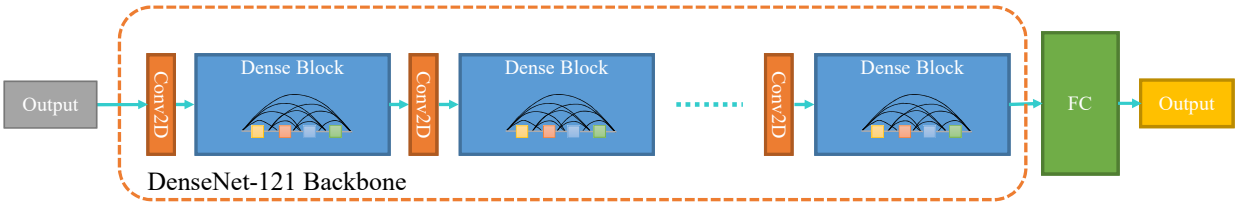


Figure 4.6: The architecture of DenseNet-121 [9].

Each dense block has four layers and the output of each layer is connected to each layer after itself.

CheXNet has been trained on frontal chest x-ray images from CXR datasets. CXR dataset is a large publicly available dataset of 14 different chest diseases. The frontal chest x-ray is indeed different from prostate MRI, however, the basic features extracted by the first layers of CheXNet are very informative. We have demonstrated that these features are useful even when they are applied to other radiology scans [78]. Fig. 4.7 depicts the architecture of our model.

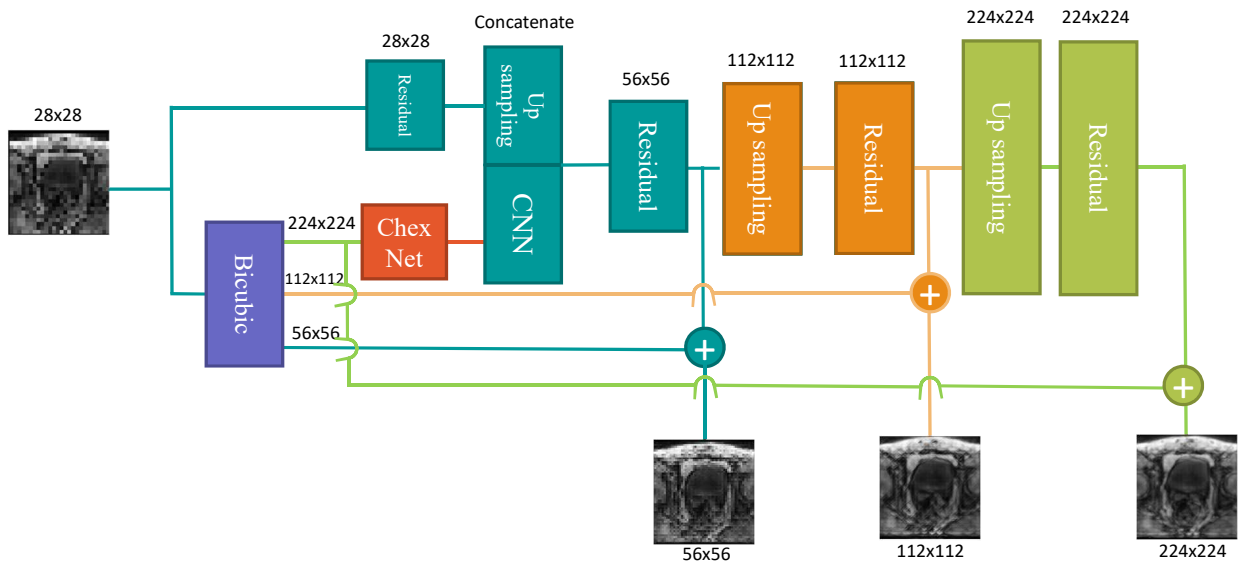


Figure 4.7: The architecture of the model for prostate MRI SR.

LR image is up-sampled with different scales for residual learning and also feature extraction using ChexNet. The first 139 layers of ChexNet with a total number of parameters of 1,444,928 are used in the proposed model. The architecture of the residual block is illustrated in Fig. 3.5. The extracted features are concatenated by a learnable CNN feature map and passed to a residual block. Each up-sampling block is followed by a residual block and the output of this block creates high-frequency details of the image in each scale. This high-frequency information is then added to a bicubic interpolation of the input image.

4.3 Dataset and Preprocessing

Prostate-Diagnosis and PROSTATEx dataset are used in this thesis [8, 79]. These datasets are publicly available on Cancer Imaging Archive [1]. These two datasets contain multi-slice MRIs of patients all diagnosed with prostate cancer, as illustrated in Fig 4.8.

Where n is the number of slices in each MRI. Each slice is a 16-bit black and white DICOM image. WL and WW are window level and window width, respectively. By changing these two parameters, the brightness and the contrast of the image can be adjusted. The data is

[1] https://www.cancerimagingarchive.net/

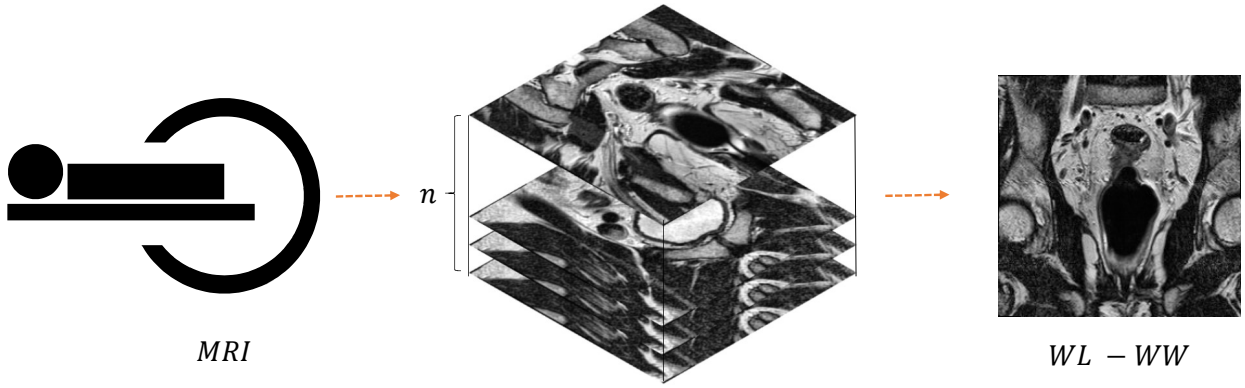


Figure 4.8: The process of obtaining dataset.

obtained from 329 patients and each patient has several scans and each scan consists of many slices. The total number of images is 329k. The scans fall into three categories of coronal, sagittal, and axial with different qualities. Some samples from the dataset are exhibited in Fig. 4.9.

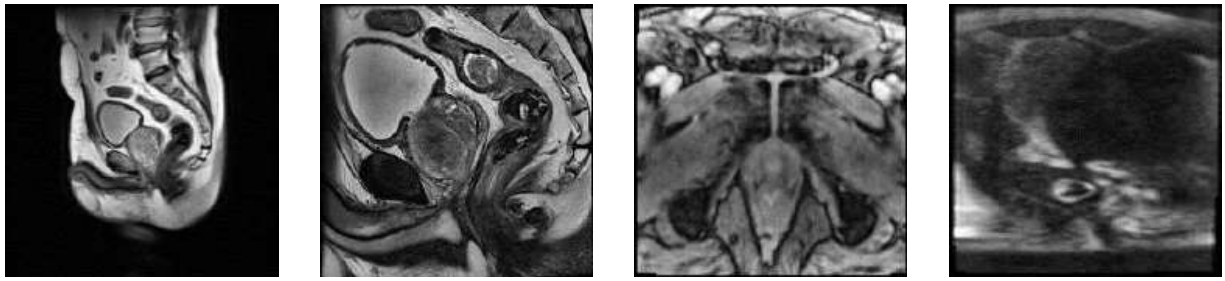
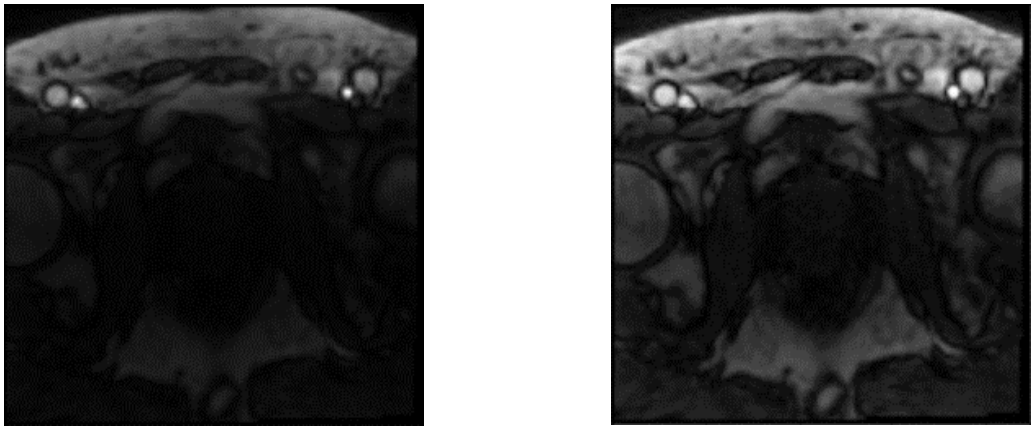


Figure 4.9: Samples from the dataset (a)wide sagittal, (b)sagittal, (c)axial, and (d)low-quality axial.

For the SR task, the data of 320 patients are used for training and the rest 9 are used for model performance evaluation.

All the samples in the dataset are in DICOM format so the first step of data preparation is reading DICOM images and resizing them to 224×224. Then, Contrast Limited Adaptive Histogram Equalization (CLAHE) is applied to each image. CLAHE is one of the most popular and powerful image enhancement algorithms [80]. Fig. 4.10 exhibits a sample from

the dataset before and after applying CLAHE.



(a) (b)

Figure 4.10: Sample image from the dataset (a) without CLAHE and (b) after applying CLAHE.

By equalizing the histogram in each region of the image, more details become visible. This preprocessing step can help the model to benefit from these details for HR image reconstruction. All the images are saved in PNG format.

In each iteration of the training, a batch of images is loaded into the memory. Then, the images are flipped horizontally with the probability of 50%. The image is down-sampled with scales of 2, 4, and 8. Finally, the format of the images changes from uint8 to float.

PROSTATEx dataset contains metadata for each patient. One of the columns in the metadata is called ClinSig. ClinSig represents whether the scan is a clinically significant finding or not. When biopsy Gleason Score is 7 or higher, this identifier will be True. This label is used as the target label for training the classifier. 4K random images are selected from the dataset with their labels. 80% are used for training and the rest 20% for testing. A problem here is dataset imbalance. Only 20% of the images are labeled as one. So in order to overcome this issue, an equal number of samples of each class is selected randomly from the dataset, instead of just a random choice regardless of the label.

4.4 Results and discussions

The model is trained with a batch size of 32. In order to avoid mode collapse and prevent the discriminator from outperforming the generator, the discriminator is trained on real samples and fake images with a batch size of 8. Other training parameters are the same as in Table 3.1.

4.4.1 Similarity Assessment

To evaluate the performance of the model, the following similarity assessment metrics have been used:

- PSNR
- SSIM
- MS-SSIM

These metrics evaluate the similarity of the super-resolved image with the ground truth. While PSNR is a pixel-wise metric, SSIM and MS-SSIM are evaluating the similarity of the distribution of the pixel values. It has been shown that these metrics fail to reflect the perceptual similarity of the output [64]. As a result, Sood *et al.* employed Mean Opinion Score (MOS) [64].

There are some important drawbacks to using MOS. First, it is extremely unlikely to reproduce the results, accurately. Hence, a comparison between models using MOS is almost impossible. Second, individuals participating in the process might be biased based on the questioner. More importantly, regarding the current performance of the DL models, the output of the SR model is usually passed to a DL-based classifier for automated diagnosis.

Regarding the problems with each aforementioned metric, a Task-Specific Similarity Assessment (TSSA) metric is proposed. TSSA evaluates the similarity of super-resolved images and the ground truth from the perspective of a DL model by investigating the impact of the SR process on the performance of the DL model. TSSA can be formulated as:

$$TSSA = \frac{L_{gt}}{L_{SR}} \tag{4.1}$$

Where L_{gt} and L_{SR} are the test loss of the classifier using ground truth and super-resolved images, respectively. Various functions can be used as L for computing TSSA. Here the accuracy score is chosen to reflect the performance of the classifier after applying SR. TSSA is usually ranged from 0 to 1. When SR has no impact on the performance of the model, TSSA will be equal to 1. The more negative impact SR has on the performance of the classifier, the closer TSSA gets to 0. In a rare condition that SR improves the accuracy of the model, TSSA will become greater than 1. In this problem, a classifier for ClinSig classification is used. The classifier has 4 convolutional layers with 64 filters and the stride of 2, followed by the output layer with the activation function of sigmoid. This CNN is trained with real 224×224 prostate scans. Then the super-resolved images are used for evaluating the TSSA, as illustrated in Fig. 4.11.

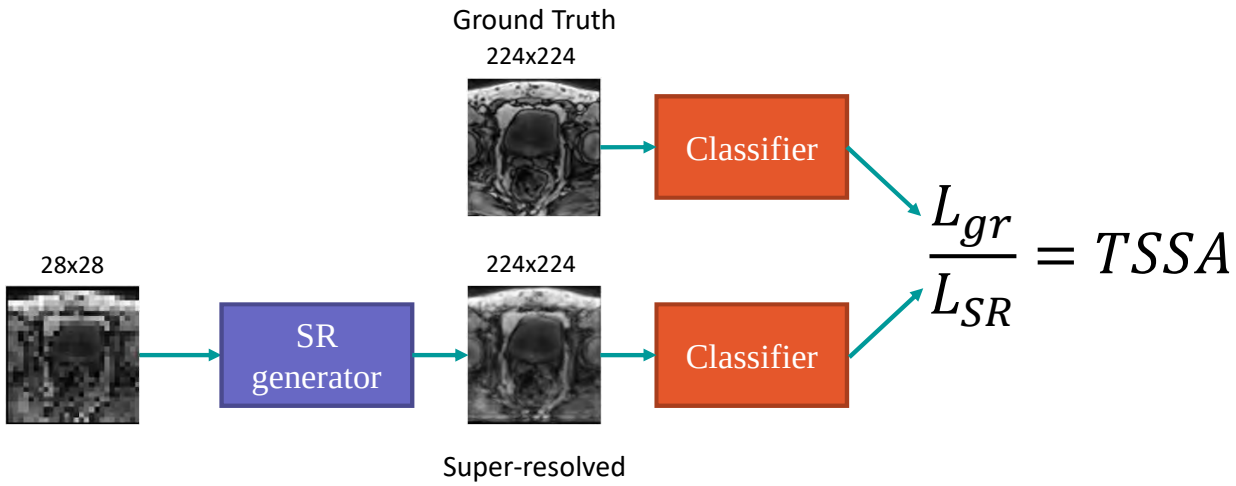


Figure 4.11: TSSA calculation for a SR model.

4.4.2 Experimental Results

After training the model, the test set is used for performance evaluation. Fig. 4.12 shows the output of the model in different stages.

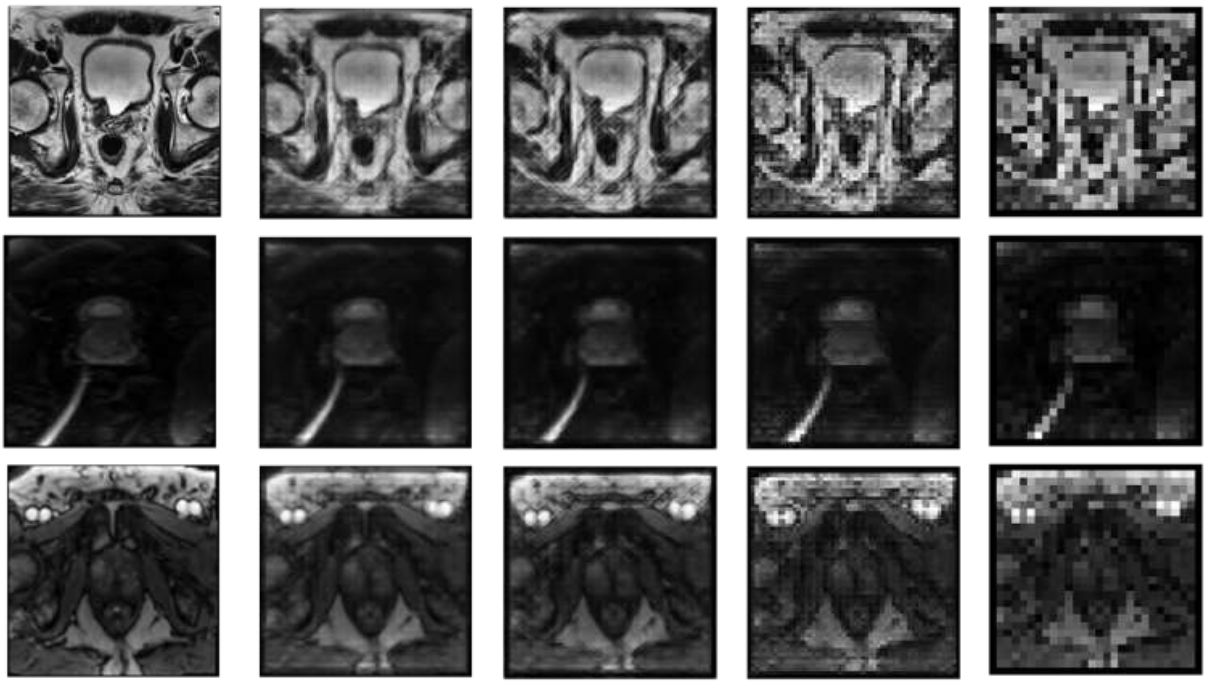


Figure 4.12: (a) Ground truth, SR out put of the proposed model with the scale of (b) 8×, (c)4×, (d)2×, and (e)LR.

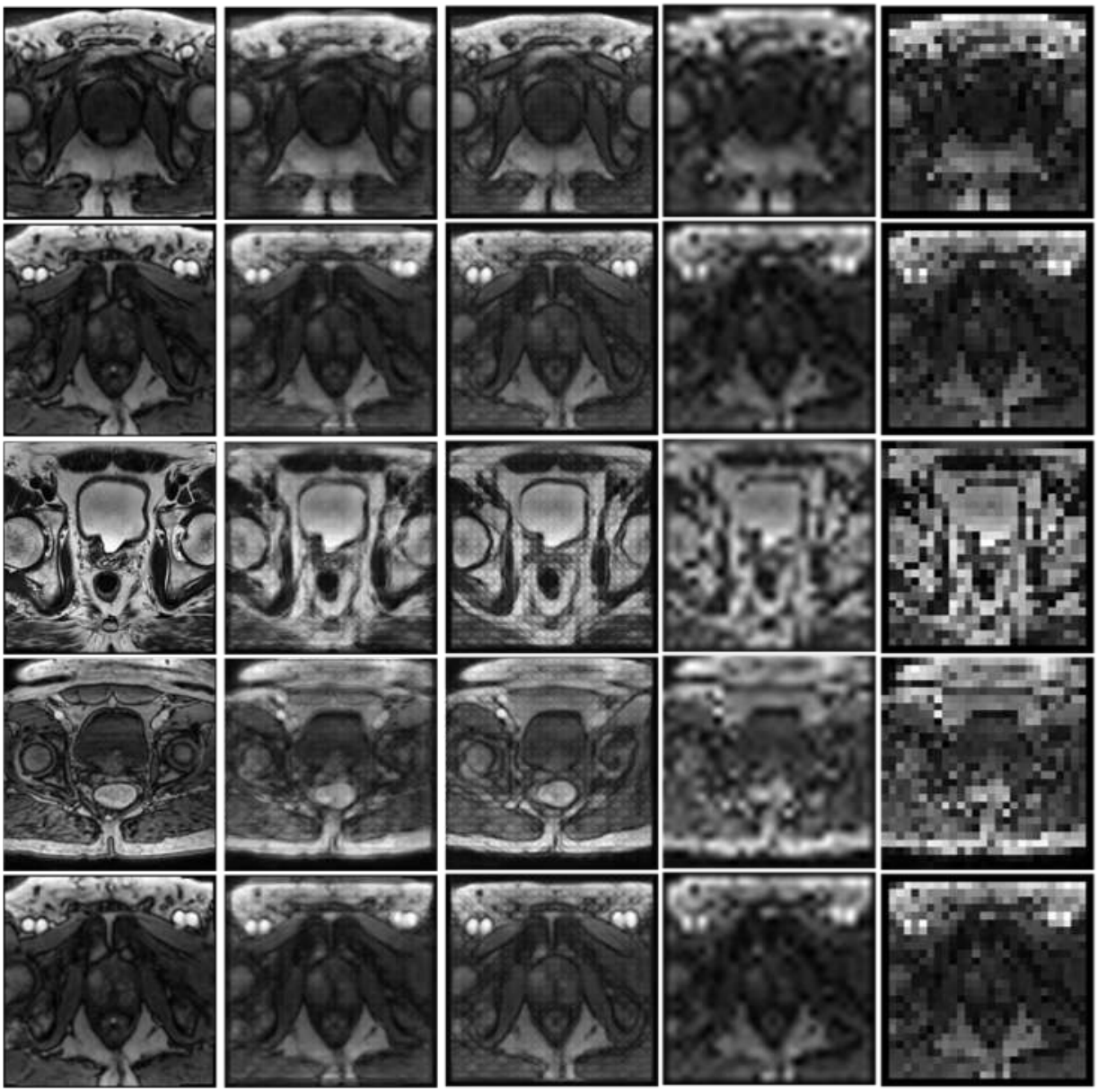
The LR image is up-sampled for different scales. Each output is connected to the discriminator. As a result, the discriminator is classifying fake and real images based on not only the ×8 images but also other scales as well.

The proposed model is compared with the state-of-the-art approach which is an SRGAN [81]. Fig. 4.13 displays a visual comparison between the outputs of the models.

To compare the models quantitatively, PSNR, SSIM, and MS-SSIM are evaluated and presented in Table 4.1.

Table 4.1: Comparison of the performance of different prostate SR models for 8× SR.

Model	**PSNR**	**SSIM**	**MS-SSIM**
Bicubic	17.92	0.50	0.69
SRGAN [64]	18.73	0.47	0.64
Proposed model	**19.77**	**0.60**	**0.79**



(a) (b) (c) (d) (e)

Figure 4.13: (a)Ground truth, (b)proposed model, (c)SRGAN, (d)bicubic, and (e)LR.

The proposed model surpassed both bicubic interpolation and the state-of-the-art SR model. Since the proposed model is benefiting from multi-scale gradient architecture, it can perform notably better in high-scale SR. However, in lower scales, the performance of the model is almost the same as other methods in terms of PSNR and SSIM as presented in Table 4.2.

Table 4.2: Comparison of the performance of different prostate SR models for 4× SR.

Model	**PSNR**	**SSIM**
SRResNet [64]	21.03	0.70
SRGAN [64]	21.27	0.66
Proposed Model	21.09	0.74

Moreover, because of the CapsNet in the discriminator, the geometrical relationships between features are more accurate in our model. Another way to compare the proposed system with the state-of-the-art is TSSA, as summarized in Table 4.3.

Table 4.3: Comparison of the performance of different prostate SR models.

Model	**Loss**	**Accuracy**	**TSSA**
Classifier	3.73	86%	-
SRGAN [64]	9.98	71%	0.82
Proposed model	7.25	79%	0.88

The results show that the proposed model achieved higher TSSA. It means that from a perspective of a deep learning-based architecture, our MSG-CapsGAN reconstructs the medical details more accurately. As far as the complexity of the models is concerned, the number of parameters are compared in Table 4.4.

The total number of parameters in the generator of the proposed model is 8% less than SRGAN. This difference is 63% in trainable parameters. The proposed model is benefiting from non-trainable embedded CheXNet for feature exaction, as a result, notably less number of trainable parameters is required for efficient SR. The proposed discriminator has 32% fewer parameters. CapsNet and MSG-GAN architecture are the main reasons for the drop in the

Table 4.4: Number of parameters in the proposed model and the state-of-the-art.

Model	Generator			Discriminator		
	trainable	non-trainable	total	trainable	non-trainable	total
SRGAN [64]	2,576,705	4,224	2,580,929	5,413,953	3,712	5,413,953
Proposed model	927,619	1,445,696	2,373,315	3,676,065	144	3,676,209

number of parameters. CapsNet represents the features by a vector so it can outperform the CNN with less number of layers [30]. MSG-GAN establishes many intermittent connections and proved the model with super-resolved images with all scales [77].

4.5 Summary

Cancer is one of the most common reasons for death worldwide. Effective treatment plans can significantly increase the survival rate of the patients and increase their life expectancy and quality. One of the most important factors in the success of a treatment plan is early diagnosis. Because of the nature of cancer, the earlier the disease is diagnosed, the more successful the treatment can be. Various approaches for cancer detection are currently used. One of the most common methods is medical imaging. Using various technologies such as MRI and ultrasound, an image of internal tissue is taken and the doctor checks the image for any abnormality and signs of cancer. In the early stage, these abnormalities are usually very small so because of the low resolution of the scans and high noise level, it is impossible to detect them. One of the most common cancers is prostate cancer. The main approach for early cancer detection and treatment planning in this cancer is MRI. In order to address the aforementioned challenges for cancer detection, especially in prostate cancer, AI can help the diagnosis process by increasing the resolution of the scan and improve the quality of the image. Thus, more details are presented in the scan and the disease can be diagnosed at an earlier stage.

One of the main concerns in medical SR, especially in high-scale SR (e.g. $8\times$), is whether the important medical details essential for diagnosis can be reconstructed or not. Moreover, in each individual case, the body or disease looks different and the system should be able

to have good performance even in different scenarios. The promising performance of the proposed face SR model in Chapter 3 motivates us to address these problems using a similar approach.

In order to employ the model proposed in Chapter 3 in prostate MRI SR several modifications are performed. First, the size of the input in the models is different (i.e. 224×224 vs 128×128). Consequently, the size of all tensors changes accordingly. An increase in the size of the input leads to an increase in the number of neurons in the layer before CapsNet. Therefore, the number of parameters grows more than 3 times of the model for face SR. In order to deal with this issue, the number of filters in the last convolutional layer is reduced to half and its stride is set to 2. In this way, the size of the output of this layer is reduced to one-fourth. Second, the pre-trained feature extractor is changed. Since face images have similarities with the ImageNet dataset, VGG19 can extract informative features from them. However, prostate MRI scans are very different from the dataset used for training VGG19. Unlike the ImageNet dataset, MRI scans are black and white and the types of edges, textures, shapes, and colors in these two datasets are completely different. As a result, using VGG19 for feature extraction in the prostate MSRI SR model is not effective and beneficial. So, VGG19 is substituted with CheXNet, a deep learning model trained for chest disease detection. CheXNet is trained on millions of chest x-ray scans. Although chest x-ray is different from prostate MRI, there are many similarities between these two types of images. The informative features extricated from CheXNet are concatenated to the features extracted using trainable convolutional layers. Then, CapsNet classifies the input as real or fake.

The model is trained with a publicly available dataset. The DICOM images are converted to PNG and the pixel values are normalized. Then, CLAHE is applied to each image to enhance the quality and the contrast. After the training process, the performance of the model is compared with the state-of-the-art work using for different scales. In $4\times$ scale, the proposed model falls behind SRGAN in PSNR and outperforms it in SSIM with a small margin. However, in $8\times$ SR, the proposed model outperforms state-of-the-art work with a notable margin in PSNR, SSIM, and MS-SSIM. The reason is by increasing the scale of SR,

the performance of SRGAN drops notably due to the lack of information in LR image and image distortion in HR output. However, the proposed model handles this issue relying on MSG-GAN architecture so the drop in the performance is notably lower in the proposed model.

As explained in Chapter 3, none of the mentioned similarity metrics cannot reflect the true similarity of images completely. PSNR evaluates the pixel-wise similarity while SSIM and MS-SSIM investigate the distribution of pixel values. This is why a task-specific similarity metric can be of a great importance for comparing the performance of the models and investigating their performance. A Task-Specific Similarity Assessment (TSSA) is proposed in this chapter. This metric evaluates the impact of SR on the performance of another classifier. A classifier is trained with ground truth images for cancer detection. Then, its accuracy is calculated using super-resolved images. A better SR model damages the performance of the classifier less. The original accuracy of the classifier is 86%. By using the images obtained from the state-of-the-art work for testing, the accuracy drops to 71%. However, the proposed model only decreases the accuracy to 79%. These results confirm that because of the robustness and the architecture of the proposed design, the medical details crucial for cancer diagnosis are reconstructed more accurately, with a notable margin.

This model demonstrates great ability in super-resolving medical scans with a promising accuracy, especially in important details required for cancer diagnosis. In order to employ this model in the medical system, a patch-based approach should be implemented. Each scan is divided into multiple 224×224 patches. Then, each patch is super-resolved independently. Finally, all the output should be stitched together to generate the complete HR scan. Another possible way to further improve the proposed model is using TSSA in the loss function. As a result, the model tries to maximize the perceptual similarity of the generated image with the ground truth, especially where important medical details present.

5. Conclusions and Future work

5.1 Conclusions

One of the most useful sub-fields of super-resolution is face SR. Given an LR image of a face, the HR counterpart is demanded. However, performing SR tasks on extremely low-resolution images is very challenging due to the image distortion in the HR results. Many deep learning-based SR approaches have intended to solve this issue by using attribute domain information. However, they require more complex data and even additional networks.

In Chapter 3, a novel robust Multi-Scale Gradient capsule GAN for face SR is proposed. 16×16 face images are up-sampled to 128×128 using a residual Multi-scale gradient capsule GAN, and it benefits from the transfer learning paradigm. The network is trained with the CelebA aligned dataset, and the performance of the network is compared with similar works. A novel metric for the similarity of images is proposed as well called Feature SIMilarity (FSIM). This network surpassed the state-of-the-art in terms of PSNR, SSIM, MS-SSIM, and FSIM without using any attribute domain information. Moreover, it is demonstrated that the proposed network is more robust to the pose variation of input images. Since no facial attribute domain information is used in this model, this SR system is not limited to face-related applications and the training is fully unsupervised. Furthermore, the robustness of the model improves the performance of the system while facing unprecedented scenarios. Relying on these two advantages, the proposed model can be considered as a general SR system.

In Chapter 4, the proposed MSG-CapsGAN is employed for prostate SR. Prostate cancer is the second most common cancer worldwide. The ability to perform SR on prostate MRI can significantly increase the chance of early cancer diagnosis. As a result, the therapy can be

started sooner and the chance of treatment success can be improved. The feature extractor in this model is CheXNet. CheXNet is trained for lung disease detection so it can extract many informative and useful features from radiography scans. The input image is 28×28 and the super-resolved output is 224×224. The results are compared with the related works and the proposed model outperforms the state-of-the-art model. Moreover, a new task-specific similarity assessment approach is introduced. This metric reflects the negative impact of the SR algorithm on the performance of a cancer detection system. While SRGAN can generate realistic outputs, it does not mean that it preserves the medical details crucial for cancer detection. In Chapter 3, it has been demonstrated that some facial expressions were not preserved in the SR process, due to the extremely low resolution of the input image. However, these details are very important in medical scans. The robustness of the proposed model and the CapsNet in the discriminator are the main reasons behind the supremacy of the proposed architecture in reconstructing important medical details.

5.2 Future work

State-of-the-art deep learning models for SR achieved promising results with the cost of high computational complexity. The implementation of these models on mobile devices for everyday applications is not practical. As a result, the necessity of designing light deep learning models is unquestionable. By developing a small model, the state-of-the-art performance cannot be reached. So the solution is to train the original model with a powerful machine and then reduce the number of parameters in the trained model. It is expected that various efficient approaches for compacting the current gigantic models will be proposed in the future.

One general issue with deep learning models is the complexity of their behavior. As a result, it is almost impossible to fully comprehend how a deep learning model does its job. The focus is usually on the performance rather than the process. However, by taking a look inside this black box and having a deep understanding of the way that model represents the data, we can improve our models significantly.

By increasing the scale in SR, the quality of the super-resolved image drops notably.

The reason is the lack of information in the LR domain and image distortion in the HR domain. Some concepts such as progressive training or Multi-scale gradient GAN has been employed by many studies to overcome this issue. However, more efficient and effective training concepts and architectures are required in order to perform accurate SR with a scale of 16 or 32. Regarding the great capability of deep learning algorithms and the numerous potential applications for high-scale SR, more research should be done on this topic. one way to address this issue is ensemble learning. Each SR model has its own advantage. So why not benefiting from all of them? All of the trained SR models can generate their outputs and another CNN combines all of these outputs and make the best super-resolved image. The CNN can be trained with the combination of GAN loss, perceptual loss, and TSSA, to make sure the output image is realistic, accurate, and useful for other DL models.

References

[1] Wikipedia contributors, "Bicubic interpolation — Wikipedia, the free encyclopedia," 2021, [Online; accessed 9-April-2021]. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Bicubic_interpolation&oldid=1005441439

[2] W. Yang, X. Zhang, Y. Tian, W. Wang, J.-H. Xue, and Q. Liao, "Deep learning for single image super-resolution: A brief review," *IEEE Transactions on Multimedia*, vol. 21, no. 12, pp. 3106–3121, 2019.

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

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

[5] D. R. Brenner, H. K. Weir, A. A. Demers, L. F. Ellison, C. Louzado, A. Shaw, D. Turner, R. R. Woods, and L. M. Smith, "Projected estimates of cancer in canada in 2020," *Cmaj*, vol. 192, no. 9, pp. E199–E205, 2020.

[6] J. Liau, D. Goldberg, and H. Arif-Tiwari, "Prostate cancer detection and diagnosis: role of ultrasound with mri correlates," *Current Radiology Reports*, vol. 7, no. 3, p. 7, 2019.

[7] "Normal prostate (mri) — radiology case — radiopaedia.org." [Online]. Available: https://radiopaedia.org/cases/normal-prostate-mri-1

[8] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, and H. Huisman, "Prostatex challenge data," *The Cancer Imaging Archive*, 2017.

[9] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya *et al.*, "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *arXiv preprint arXiv:1711.05225*, 2017.

[10] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1153–1160, 1981.

[11] C. E. Duchon, "Lanczos filtering in one and two dimensions," *Journal of Applied Meteorology and Climatology*, vol. 18, no. 8, pp. 1016–1022, 1979.

[12] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Computer graphics and Applications*, vol. 22, no. 2, pp. 56–65, 2002.

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

[14] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2861–2873, 2010.

[15] S. Schulter, C. Leistner, and H. Bischof, "Fast and accurate image upscaling with super-resolution forests," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3791–3799.

[16] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

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

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

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

[20] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3147–3155.

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

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

[23] R. Kalarot, T. Li, and F. Porikli, "Component attention guided face super-resolution network: Cagface," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 370–380.

[24] M. M. Majdabadi and S.-B. Ko, "Capsule gan for robust face super resolution," *Multimedia Tools and Applications*, vol. 79, no. 41, pp. 31 205–31 218, 2020.

[25] H. A. Song and S.-Y. Lee, "Hierarchical representation using nmf," in *International conference on neural information processing*. Springer, 2013, pp. 466–473.

[26] N. Rochester, J. Holland, L. Haibt, and W. Duda, "Tests on a cell assembly theory of the action of the brain, using a large digital computer," *IRE Transactions on information Theory*, vol. 2, no. 3, pp. 80–93, 1956.

[27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[28] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[29] J. L. Elman, "Finding structure in time," *Cognitive science*, vol. 14, no. 2, pp. 179–211, 1990.

[30] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in neural information processing systems*, 2017, pp. 3856–3866.

[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[32] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[33] C.-H. Lee, K. Zhang, H.-C. Lee, C.-W. Cheng, and W. Hsu, "Attribute augmented convolutional neural network for face hallucination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 721–729.

[34] S. S. Rajput and K. Arya, "A robust facial image super-resolution model via mirror-patch based neighbor representation," *Multimedia Tools and Applications*, vol. 78, no. 18, pp. 25 407–25 426, 2019.

[35] C. Liu, H.-Y. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *International Journal of Computer Vision*, vol. 75, no. 1, pp. 115–134, 2007.

[36] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 35, no. 3, pp. 425–434, 2005.

[37] S. Kolouri and G. K. Rohde, "Transport-based single frame super resolution of very low resolution face images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4876–4884.

[38] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *European conference on computer vision*. Springer, 2016, pp. 318–333.

[39] X. Yu and f. Porikli, "Face hallucination with tiny unaligned images by transformative discriminative neural networks," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[40] X. Yu and F. Porikli, "Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3760–3768.

[41] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," in *European conference on computer vision*. Springer, 2016, pp. 614–630.

[42] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

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

[44] M. M. Majdabadi and S.-B. Ko, "Msg-capsgan: Multi-scale gradient capsule gan for face super resolution," in *International Conference on Electronics, Information, and Communication (ICEIC)*. IEEE, 2020, pp. 1–3.

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

[46] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *European conference on computer vision*. Springer, 2016, pp. 391–407.

[47] G. Montúfar, R. Pascanu, K. Cho, and Y. Bengio, "On the number of linear regions of deep neural networks," *arXiv preprint arXiv:1402.1869*, 2014.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

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

[50] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Learning face hallucination in the wild," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.

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

[52] X. Xu, D. Sun, J. Pan, Y. Zhang, H. Pfister, and M.-H. Yang, "Learning to super-resolve blurry face and text images," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 251–260.

[53] X. Yu, B. Fernando, R. Hartley, and F. Porikli, "Super-resolving very low-resolution face images with supplementary attributes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 908–917.

[54] M. Li, Y. Sun, Z. Zhang, H. Xie, and J. Yu, "Deep learning face hallucination via attributes transfer and enhancement," in *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 604–609.

[55] Z.-S. Liu, W.-C. Siu, and Y.-L. Chan, "Reference based face super-resolution," *IEEE Access*, vol. 7, pp. 129 112–129 126, 2019.

[56] C. Wang, Z. Zhong, J. Jiang, D. Zhai, and X. Liu, "Parsing map guided multi-scale attention network for face hallucination," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2518–2522.

[57] S. Ge, S. Zhao, C. Li, Y. Zhang, and J. Li, "Efficient low-resolution face recognition via bridge distillation," *IEEE Transactions on Image Processing*, vol. 29, pp. 6898–6908, 2020.

[58] F. Rousseau, A. D. N. Initiative *et al.*, "A non-local approach for image super-resolution using intermodality priors," *Medical image analysis*, vol. 14, no. 4, pp. 594–605, 2010.

[59] R. R. Peeters, P. Kornprobst, M. Nikolova, S. Sunaert, T. Vieville, G. Malandain, R. Deriche, O. Faugeras, M. Ng, and P. Van Hecke, "The use of super-resolution techniques to reduce slice thickness in functional mri," *International Journal of Imaging Systems and Technology*, vol. 14, no. 3, pp. 131–138, 2004.

[60] X. Yang, S. Zhan, C. Hu, Z. Liang, and D. Xie, "Super-resolution of medical image using representation learning," in *8th International Conference on Wireless Communications & Signal Processing (WCSP)*. IEEE, 2016, pp. 1–6.

[61] J. Park, D. Hwang, K. Y. Kim, S. K. Kang, Y. K. Kim, and J. S. Lee, "Computed tomography super-resolution using deep convolutional neural network," *Physics in Medicine & Biology*, vol. 63, no. 14, p. 145011, 2018.

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

[63] Y. Chen, Y. Xie, Z. Zhou, F. Shi, A. G. Christodoulou, and D. Li, "Brain mri super resolution using 3d deep densely connected neural networks," in *IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 739–742.

[64] R. Sood, B. Topiwala, K. Choutagunta, R. Sood, and M. Rusu, "An application of generative adversarial networks for super resolution medical imaging," in *17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018, pp. 326–331.

[65] Z. Li, Y. Wang, and J. Yu, "Reconstruction of thin-slice medical images using generative adversarial network," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2017, pp. 325–333.

[66] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 3856–3866. [Online]. Available: http://papers.nips.cc/paper/6975-dynamic-routing-between-capsules.pdf

[67] G. A. Bryant, "Animal signals and emotion in music: Coordinating affect across groups," *Frontiers in Psychology*, vol. 4, p. 990, 2013.

[68] A. Karnewar, O. Wang, and R. S. Iyengar, "MSG-GAN: multi-scale gradient GAN for stable image synthesis," *CoRR*, vol. abs/1903.06048, 2019. [Online]. Available: http://arxiv.org/abs/1903.06048

[69] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[70] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2472–2481.

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

[72] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[73] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image

quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.

[74] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[75] I. Reda, A. Khalil, M. Elmogy, A. Abou El-Fetouh, A. Shalaby, M. Abou El-Ghar, A. Elmaghraby, M. Ghazal, and A. El-Baz, "Deep learning role in early diagnosis of prostate cancer," *Technology in cancer research & treatment*, vol. 17, p. 1533034618775530, 2018.

[76] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[77] A. Karnewar and O. Wang, "Msg-gan: Multi-scale gradients for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7799–7808.

[78] A. Haghanifar, M. M. Majdabadi, and S.-B. Ko, "Paxnet: Dental caries detection in panoramic x-ray using ensemble transfer learning and capsule classifier," *arXiv preprint arXiv:2012.13666*, 2020.

[79] Bloch, B. Nicolas, Jain, Ashali, and C. C. Jaffe, "Data from prostate-diagnosis," *The Cancer Imaging Archive*, 2015.

[80] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer vision, graphics, and image processing*, vol. 39, no. 3, pp. 355–368, 1987.

[81] R. Sood and M. Rusu, "Anisotropic super resolution in prostate mri using super resolution generative adversarial networks," in *IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE, 2019, pp. 1688–1691.