Face Frontalization using Hybrid Attention and Relative Average Discriminator based Generative Adversarial Network

Seifedin Hayredin Mohammed

A Thesis submitted to the Department of Computer Science and Engineering,

School of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of Master's in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

June, 2024

Adama, Ethiopia

Face Frontalization using Hybrid Attention and Relative Average Discriminator based Generative Adversarial Network

Seifedin Hayredin Mohammed

Advisor: Worku Jifara Sori (Ph.D.)

A Thesis Submitted to the Department of Computer Science and Engineering,

School of Electrical Engineering and Computing

Presented in Partial Fulfillment of the Requirement for the Degree of Master's in Computer Science and Engineering

Office of Graduate Studies

Adama Science and Technology University

June, 2024

Adama, Ethiopia

DECLARATION

I hereby declare that this Master Thesis entitled "**Face Frontalization using Hybrid Attention and Relative Average Discriminator based Generative Adversarial Network"** is my original work. That is, it has not been submitted for the award of any academic degree, diploma or certificate in any other university. All sources of materials that are used for this thesis have been duly acknowledged through citation

Name of student	Signature	Date

RECOMMENDATION OF ADVISOR

I, the advisor of this thesis, hereby certify that I have read the revised version of the thesis entitled “**Face Frontalization using Hybrid Attention and Relative Average Discriminator based Generative Adversarial Network”** prepared under my guidance by **Seifedin Hayredin Mohammed** submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science and Engineering. Therefore, I recommend the submission of the revised version of the thesis to the department following applicable procedures.

Major Advisor	Signature	Date

APPROVAL SHEET

I, the advisor of the thesis entitled "**Face Frontalization using Hybrid Attention and Relative Average Discriminator based Generative Adversarial Network"** and developed by **Seifedin Hayredin Mohammed**, hereby certify that the recommendation and suggestions made by the board of examiners are appropriately incorporated into the final version of the thesis.

Major Advisor	Signature	Date

APPROVAL OF BOARD OF REVIEWERS

We, the undersigned, members of the Board of Examiners of the thesis by **Seifedin Hayredin Mohammed** have read and evaluated the thesis entitled "**Face Frontalization using Hybrid Attention and Relative Average Discriminator based Generative Adversarial Network**" and examined the candidate during open defense. This is, therefore, to certify that the thesis is accepted for partial fulfillment of the requirement of the degree of Master of Science in Computer Science and Engineering.

Chairperson	Signature	Date
Internal Examiner	Signature	Date
External Examiner	Signature	Date

Finally, approval and acceptance of the thesis is contingent upon submission of its final copy to the Office of Postgraduate Studies (OPGS) through the Department Graduate Council (DGC) and School Graduate Committee (SGC).

Department Head	Signature	Date
School Dean	Signature	Date
Office of Postgraduate Studies, Dean	Signature	Date

ACKNOWLEDGMENT

I would like to begin by expressing my deepest gratitude to **Allah**. Throughout my studies, I am incredibly thankful for the blessings of good health and resilience. These gifts allowed me to persevere and reach this point.

I would like to extend my sincere thanks to **Dr. Worku Jifara (Ph.D.),** my advisor, whose guidance was instrumental in the completion of this work. His unwavering guidance, support, and insightful suggestions proved to be an invaluable resource throughout my entire research journey. I would also like to extend my gratitude to **Dr. Mesfin Abeba (Ph.D.)** for his persistent supervision and valuable recommendations throughout the duration of this study.

Finally, I would like to express my sincere gratitude to the members of the artificial intelligence SIG, my family, and friends.

TABLE OF CONTENTS

ACKNOWLEDGMENT i

LIST OF TABLES vii

LIST OF FIGURES viii

LIST OF ACRONYMS ix

ABSTRACT x

CHAPTER ONE 1

1. INTRODUCTION 1

1.1 Background of the Study 1

1.2 Motivation of the Study 4

1.3 Statement of the Problem 4

1.4 Research Questions 5

1.5 Objectives of the Study 5

1.5.1 General Objective 5

1.5.2 Specific Objectives 6

1.6 Significances of the Study 6

1.7 Scopes and Limitations of the Study 7

1.7.1 Scopes of the Study 7

1.7.2 Limitations of the Study 7

1.8 Contributions of the Study 7

1.9 Organizations of the Study 8

CHAPTER TWO 9

2. LITERATURE REVIEW 9

2.1 Overview of Face Frontalization 9

2.2 Face Frontalization Techniques 9

2.2.1 Traditional Face Frontalization Techniques ... 9
2.2.2 Techniques for Face Frontalization using Deep Learning ... 10
2.3 Generative Adversarial Network (GAN) ... 12
2.4 Generative Adversarial Networks Architectures ... 13
2.4.1 Conditional GAN ... 13
2.4.2 Least Squares GAN ... 14
2.4.3 CycleGAN ... 15
2.4.4 Deep Convolutional GAN (DCGAN) ... 16
2.4.5 Relativistic Discriminator ... 17
2.4.6 Relativistic Average Discriminator ... 17
2.5 U-Net ... 18
2.6 Translation of Image to Image (I2I) ... 18
2.7 Face Synthesis ... 19
2.8 Face Recognition ... 20
2.9 Facial Attribute Manipulation ... 20
2.10 Related Works ... 21
2.11 Summary of Related Works ... 23
CHAPTER THREE ... 26
3. RESEARCH METHODOLOGY ... 26
3.1 Overview of Chapter ... 26
3.2 Datasets ... 27
3.2.1 CAS-PEAL-R1 Dataset ... 27
3.2.2 Multi-PIE Dataset ... 29
3.3 Pre-processing Techniques ... 30
3.3.1 Data Augmentation ... 30

3.3.2 Normalization 30
3.3.3 Resize 30
3.4 Development Tools 31
3.4.1 Hardware Tools 31
3.4.2 Software Tools 31
3.5 Baseline Works 32
3.6 Feature Extraction Network 32
3.7 Evaluation Metrics 33
3.7.1 Rank-1 Recognition Rate 33
3.7.2 PSNR (Peak Signal to Noise Ratio) 33
3.7.3 SSIM (Structural Similarity Index Measure) 34
CHAPTER FOUR 35
4. PROPOSED ARCHITECTURE 35
4.1 Overview of Chapter 35
4.2 Model Architecture 35
4.3 Generator Architecture 38
4.3.1 Self-attention 41
4.3.2 Channel Attention 42
4.4 Discriminator Network 43
4.4.1 Rationality for using Relative Average Discriminator 44
4.5 Learning Functions of the Model 47
4.5.1 L_2 Loss 47
4.5.2 Multi-scale Pixel-wise Loss 47
4.5.3 Total Variation Regularization Loss 48
4.6 Training Pseudo Code 48

CHAPTER FIVE 50
5. IMPLEMENTATION DETAILS 50
5.1 Chapter Overview 50
5.2 Working Environment 50
5.3 Environmental Setup 50
5.4 Implementation of Proposed Model 51
5.4.1 Implementation of Self-attention 54
5.4.2 Implementation of Channel Attention 55
5.4.3 Implementation of Relative Average Discriminator 55
5.5 Hyperparameter Configuration 56
5.6 Experimental Class 57
CHAPTER SIX 60
6. RESULTS AND DISCUSSION 60
6.1 Overview of Chapter 60
6.2 Experimental Results 60
6.3 Results based on Evaluation Metrics 60
6.3.1 Evaluation of Experiments on Multi-PIE dataset 61
6.3.2 Evaluation of Experiments on CAS-PEAL-R1 dataset 62
6.4 Frontalized Image Results of Experiments 64
6.4.1 Frontalized Results of Experiments for CAS-PEAL-R1 dataset 64
6.4.2 Frontalized Results of Experiments for Multi-PIE dataset 65
6.4.3 DMA-GAN + SA (FFRAD-GAN-1) 66
6.4.4 DMA-GAN + SA + CA (FFRAD-GAN-2) 66
6.4.5 DMA-GAN + SA + RAD (FFRAD-GAN-3) 67
6.4.6 DMA-GAN + SA + CA +RAD (FFRAD-GAN-4) 68

6.5 Sample Training Log for FFRAD-GAN 69
6.6 Research Question and Answer Discussion 72
CHAPTER SEVEN 73
7. CONCLUSION AND FUTURE WORKS 73
7.1 Conclusion 73
7.2 Future Works 74
SPECIAL ACKNOWLEDGMENT 75
REFERENCES 76

LIST OF TABLES

Table 2. 1: Summary of related work 23
Table 5. 1: Content of encoder block in generator architecture 52
Table 5. 2: Content of bottleneck layer 52
Table 5. 3: Content of decoder block in generator architecture 53
Table 5. 4: Content of relative average discriminator architecture 54
Table 5. 5: Hyperparameter configuration 56
Table 5. 6: List of experiment class 58
Table 6. 1: PSNR and SSIM evaluation on Multi-PIE dataset 61
Table 6. 2: Rank-1 recognition rate evaluation on Multi-PIE dataset 61
Table 6. 3: PSNR and SSIM evaluation on CAS-PEAL-R1 dataset 62
Table 6. 4: Rank-1 recognition rate evaluation on CAS-PEAL-R1 dataset 62

LIST OF FIGURES

Figure 2. 1 Architecture of GAN 13
Figure 2. 2 CGAN architecture 14
Figure 2. 3: cycleGAN Architecture 16
Figure 2. 4: DCGAN Architecture 16
Figure 2. 5 U-net architecture 18
Figure 2. 6 Translation of image 19
Figure 3. 1 Overall process 26
Figure 3. 2 CAS- PEAL- R1 dataset samples 28
Figure 3. 3 Multi-PIE dataset samples 29
Figure 3. 4 ResNet18 architecture 33
Figure 4. 1 Overall architecture of the proposed model 37
Figure 4. 2 Generator architecture 40
Figure 4. 3 Self-attention based GAN block. 42
Figure 4. 4 Channel attention block. 43
Figure 4. 5 Architecture of discriminator 46
Figure 6. 1: CAS-PEAL-R1 frontalized facial images in different experiment classes. 64
Figure 6. 2: Multi-PIE frontalized facial images in different experiment classes 65
Figure 6. 3: Results of DMA-GAN + SA 66
Figure 6. 4: Results of DMA-GAN + SA + CA 67
Figure 6. 5: Results of DMA-GAN + SA +RAD 67
Figure 6. 6 : Results of DMA-GAN + SA + CA + RAD (FFRAD-GAN) 68
Figure 6. 7: Discriminator loss during training on CAS-PEAL-R1 dataset 69
Figure 6. 8: Generator loss during training on CAS-PEAL-R1 dataset 69
Figure 6. 9: Multi scale pixel-wise loss during training on CAS-PEAL-R1 dataset 70
Figure 6. 10: Discriminator loss during training on Multi-PIE dataset 70
Figure 6. 11: Generator loss during training on Multi-PIE dataset 71
Figure 6. 12: Multi scale pixel-wise loss during training on Multi-PIE dataset 71

LIST OF ACRONYMS

2D	Two Dimensional
3D	Three Dimensional
CAS	Chines Academy of Science
CA	Channel Attention
CNN	Convolutional Neural Network
FFM	Feature Fusion Module
FFRAD	Face Frontalization using Relative Average Discriminator
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
MSE	Mean Squared Error
Multi-PIE	Multi-Pose, Illumination and Expression
PEAL	Pose, Expression, Accessory and Lighting
PSNR	Peak Signal to Noise Ratio
RAD	Relative Average Discriminator
RAM	Random Access Memory
SA	Self Attention
SARA	Self Attention and Relative Average
SSIM	Structural Similarity Index Measure

ABSTRACT

Facial recognition technology is becoming more widely used in fields such as security, marketing, healthcare, and entertainment. However, in real-world scenario, the acquired facial images are often non-frontal due to acquisition angles and environmental conditions, results in variability of human faces due to pose and illumination. Side-view faces, due to varying acquisition angles and environmental factors, often lead to reduced facial recognition accuracy. Therefore, frontalizing these images to a frontal view is crucial. Due to advancement of Generative Adversarial Networks (GANs), face frontalization research has shown improvement and enhances performance of face recognition technology. However, these methods still struggle with large pose angles above 45° and do not recover a facial attribute like eyeglass, hair and skin texture. In this paper we proposed face frontalization using combination of attention and relative average discriminator on GAN (FFRAD-GAN) designed to improve frontalization of non-frontal faces, especially under large pose angles above 45°. In the generator network our proposed approach utilized U-net with combination of attention (self-attention and channel-attention) mechanism. We utilized channel attention in the encoder section, self-attention in the decoder section of the U-net based generator, and relative average discriminator in the discriminator block of the network. The proposed approach FFRAD-GAN used CAS-PEAL-R1 and Multi-PIE datasets for training and testing, and compared its performance with the baseline work. The frontalized facial image quality was evaluated using Rank1-recognition rate, peak signal to noise ratio (PSNR) and Structural similarity index (SSIM). Based on Rank1-recognition rate our FFRAD-GAN model improves the baseline work DMA-GAN from 98.98 to 99.24 on 15°, 96.59 to 97.19 on 30° and 93.18 to 94.10 on 45°. Based on SSIM and PSNR our model also got 38.76 and 0.9969 values respectively. The results indicated that FFRAD-GAN significantly outperforms other face frontalization approaches like TP-GAN, GSP-GAN, DA-GAN and DMA-GAN in terms of frontalizing facial images with a large pose angle above 45° and producing high quality frontalized facial images. So, FFRAD-GAN have a significant impact on face recognition and surveillance systems that requires high-quality frontal facial images for performing their tasks effectively.

Keywords: Face frontalization, generative adversarial network, self-attention, channel attention, relative average discriminator.

CHAPTER ONE

1. INTRODUCTION

1.1 Background of the Study

With the rise of digital connections everywhere, applications of facial recognition technology have become increasingly prevalent, spanning across various domains such as security, marketing, healthcare, and entertainment. However, due to environmental conditions and acquisition angles, the acquired images in practical applications are typically non-frontal faces.

A major challenge for facial recognition systems is the natural variability in human faces caused by differences in pose, lighting, and expressions. These disruptions frequently result in a notable decline in accuracy of facial recognition, mainly in larger poses (angles above 45˚). Currently, this topic is being explored essentially through research pathways: approach for acquiring features which are not vary with pose (Ding & Tao, 2015,) (Ding & Tao, 2017) and techniques for transforming a facial image from any pose to its canonical frontal view, a process known as face frontalization. (Luan et al., 2020; Junho Yim et al., 2015; Huang et al., 2017; Yin et al., 2020; Zhang et al., 2021; Cen et al., 2022; Luo et al., 2022).

Face frontalization is a vital task in area of computer vision and facial recognition. It involves transformation of a given face image from an arbitrary pose or orientation to a canonical frontal view, which is essential for many applications, including face recognition, emotion analysis, and human-computer interaction. Achieving accurate and robust face frontalization has significant implications for improving the performance of these applications. To address this challenge and enhance the precision and resilience of facial recognition systems, the concept of face frontalization has emerged as a pivotal area of research and development. Many outstanding works have emerged.

Face frontalization research has been done in different techniques. Face frontalization based on 2D/3D local texture warping (Hassner et al., 2015;Asthana et al., 2011;Jeni & Cohn, 2016; X. Yin et al., 2017). (Hassner et al., 2015), discover the technique that utilizes a single, unaltered surface of 3D in order to represent the contours of each input face. 2D local texture warping techniques that work mostly in the 2D image domain are unable to leverage 3D facial geometry because of the lack of 3D information. This can limit their ability to handle non-frontal faces with depth variations.

Face frontalization based on statistical modeling(Wenchao Zhang et al., 2005; Sagonas et al., 2015; Sagonas et al., 2017). (Wenchao Zhang et al., 2005) proposes the use of a novel face-representation technique called Local Gabor Binary Pattern Histogram Sequence (LGBPHS), that eliminates needs for statistical techniques and a training process for building face models. To avoid the issue of generalizability, the image of a face is depicted as a sequence of histograms achieved by combining all local regions histograms from the binary pattern maps of Local Gabor Magnitude. Due to variability with face poses and illumination, Statistical models are ineffective to transform the image to a common frontal view.

Deep learning-based face frontalization utilizing Convolutional Neural Networks (Kang et al., 2018; Al-Azzawi et al., 2018; Z. Zhang et al., 2019). (Z. Zhang et al., 2019) introduced Frontalization of Face with Convolutional Neural Network utilizing Appearance Flow (A3F-CNN) which learns to construct a rich correlation between frontal-view and side-view faces. After the correlation is established, canonical-view faces are generated through explicitly shifting the side-view face pixels. While CNNs can learn to capture certain facial features and transformations, they struggle to handle the more complex non-rigid deformations in facial expressions and poses and are ineffective at creating new, realistic frontal perspectives of faces.

Facial frontalization employing deep adversarial learning between the generators and discriminators. Since (Goodfellow et al., 2014) presented the generative adversarial network, GAN-based research on frontalization of face has been increasing due to the great generating power of generative adversarial networks (GAN) (Luan et al., 2020; Huang et al., 2017; Yin et al., 2020; Cen et al., 2022; Luo et al., 2022). Luan et al., 2020, proposed a new method for frontalizing faces from various angles and then recognizing them. Their method uses a special type of Generative Adversarial Network (GAN) called a Geometry Structure Preserving GAN (GSP-GAN) to achieve this. In this approach, the model's generator is designed as a conventional auto-encoder, where the encoder retrieves identity details and the decoder generates a suitable frontal facial image. Additionally, a self-attention is incorporated into the discriminator to maintain natural facial structure. Yin et al., 2020, proposed Dual-Attention Generative Adversarial Network (DA-GAN) for creating photorealistic frontal views of face by considering both finer details and contextual dependencies. They presented a generator based on self-attention to combine local features with their distant relationships to produce improved feature descriptors and face-attention-based discriminator with four separate discriminators to highlight local aspects of face regions. However, generating better-quality frontal views of faces remain to be a significant challenge and also attributes of face like eyeglass, skin, and hair are not recovered well in extreme poses above 45˚.

Previous works use a standard GAN in which its discriminator provides a probability that input image is real without considering the fact that only half of the input data is real while the other half is synthetic. When the synthesized data closely resembles genuine data, the discriminator categorizes it as real, leading to a decline in the discriminator's effectiveness. Consequently, the generator's training cannot proceed effectively. The relative average discriminator, proposed by (Jolicoeur-Martineau, 2019) is one approach to addressing this problem. This improved discriminator evaluates the likelihood that a given actual data sample is more probable to be real than randomly selected fake data samples, rather than simply identifying the input data as real or synthetic. But the standard GAN-based approaches do not consider the prior information regarding the input data for the discriminator.

Following the proposal of the generative adversarial network (Goodfellow et al., 2014), research work on face frontalization based on GAN has been increasing due to the great generative capability of generative adversarial networks (GAN) and attention mechanisms are frequently employed and provide incredible results (Luan et al., 2020; Huang et al., 2017; Yin et al., 2020; Cen et al., 2022; Luo et al., 2022; Cao et al., 2023). This work proposed to produce high-quality and photo-realistic faces especially in larger poses inspired by DMA-GAN (Cao et al., 2023). We proposed self-attention and channel-attention in GAN's U-net-based generator network based on previous studies.

Inspired by RAD (Jolicoeur-Martineau, 2019) and SARA-GAN (Yuan et al., 2020), We proposed a discriminator to shift from classifying absolutely real or generated to a relative approach which compares the quality of the generated image to the average quality of real images.

1.2 Motivation of the Study

Recently, facial recognition technology has shown satisfactory results in the frontal perspective, However, the acquired facial images are mainly non-frontal view due to ambient conditions and acquiring angles. These disruptions often lead to a notable decline in performance of face recognition, especially with larger poses above 45˚. These problems can be solved by a face frontalization mechanism.

GANs have become a game-changer in frontalization of face, leading to impressive advancements in recent times. However, generating better-quality facial image becomes a challenging task and the facial attributes like eyeglass, skin, and hair are not recovered well especially in extreme poses above 45˚ (Yin et al., 2020; Luan et al., 2020; Cao et al., 2023). The ability to generate high-quality faces with their facial attributes is valuable in areas like face recognition: privacy protection, security systems, access control, and identity verification. It can serve as a method for augmentation and a preprocessing of data in face recognition.

Recent methods which proposed the face frontalization are commonly used only to work on convolutional layers. Convolutional neural networks (CNNs) excel at extracting local features within an image due to the localized connections between neurons in convolutional layers. In a convolutional layer of a neural network, each processing unit (neuron) only considers a specific area of the data it receives from the previous layer. This localized connectivity, while efficient for capturing local patterns, presents a significant limitation: it hinders the network's ability to effectively compute distant relationship across the entire image. So, to recover the facial attributes which is essential in applications of a computer vision, it is better to focus on the most relevant facial features for frontalization. While significant progress has been made in face frontalization, certain challenges remain open for further exploration.

1.3 Statement of the Problem

Frontal view facial image is a critical component of many computer vision applications, like facial recognition, augmented reality, and surveillance systems. Due to environmental constraints and limitations in capture angles, the majority of facial images acquired are side-views. This poses a significant challenge for facial recognition systems. Unlike frontal images where facial features are clearly presented, non-frontal views, particularly those with large pose angle above 45˚, often obscure critical facial attributes. This makes it extremely difficult to accurately identify individuals within these images. As a result, achieving reliable facial recognition with predominantly side-view data remains an ongoing challenge in the field. Since non-frontal and side-view facial images lose certain facial attributes and details, they have challenges for computer vision applications like facial recognition, augmented reality, and surveillance systems. Therefore, a face frontalization mechanism is necessary to preprocess non-frontal facial images before utilizing them in such applications. Previous works like DA-GAN (Yin et al., 2020), GSP-GAN (Luan et al., 2020), FI-GAN (Rong et al., 2020) and DMA-GAN (Cao et al., 2023) proposed face frontalization based on GAN and attention mechanism. However, these works do not focus on relevant facial attributes and do not recover relevant facial attributes like eyeglass, skin and hair in extreme poses above 45˚ (Yin et al., 2020; Luan et al., 2020; Cao et al., 2023).

To overcome those problems, we proposed face frontalization utilizing combination of attention and relative average based discriminator.

1.4 Research Questions

This work tried to address the research questions listed below.

RQ1: What are the effects of adding an attention mechanism and relative average discriminator regarding improving the quality, and frontalizing facial images with a large pose angle?

RQ2: What effective constraints can be employed to enhance the quality of synthesized frontal faces?

1.5 Objectives of the Study

1.5.1 General Objective

The general objective of this study is to develop a GAN based model for face frontalization using hybrid attention and relative average discriminator.

1.5.2 Specific Objectives

The specific objectives outlined below were addressed to accomplish the general objective.

- To review related works and identify current trends to come up with gaps on face frontalization for synthesizing frontal faces.
- To collect facial image datasets among publicly available face datasets which contains a range of variations in pose, age, and gender.
- To pre-process available data to make it suitable for our model.
- To design GAN based architecture by adding attention mechanisms and relative average discriminator.
- To train the model on the available side-view and frontal facial data.
- To evaluate the model's performance using different metrics of evaluation on test data.

1.6 Significances of the Study

This study has several benefits in the fields listed below.

- Facial Recognition Technology: The accuracy and reliability of facial recognition systems are greatly enhanced by face frontalization. Faces frequently appear at different angles in real-world situations; frontalization helps to normalize these variations, which facilitates the effective operation of recognition algorithms.
- Security Applications: The creation of robust face frontalization methods has a big impact on security applications like identity verification, access control, and surveillance systems. Accurate face alignment is critical for these systems to improve security and reduce false positives and negatives.
- Human-Computer Interaction: By enabling more precise and natural facial expression analysis, gaze tracking, and emotion recognition, face frontalization can improve human-computer interaction. This holds significance for applications like gaming, augmented, and virtual reality.

1.7 Scopes and Limitations of the Study

1.7.1 Scopes of the Study

The aim of this study is to generate photo-realistic frontal-view face images from side-view images while preserving their relevant facial attributes. This is achieved through the implementation of combining an attention mechanisms and relative average discriminator on GAN network, designed to improve face frontalization tasks. The proposed model is conditioned by inputting non-frontal face images. Therefore, the scope encompasses the development of GAN network with combination of an attention and relative average discriminator, followed by the evaluation of its results in comparison to previous approaches.

1.7.2 Limitations of the Study

Due to time and resource constraints, the following were not covered in this study.

- Our work does not involve face recognition. Instead, our model excels at transforming non-frontal facial images into a frontal view.
- Our study only conducted by considering the pose angles of image up to 90°. The pose angles above 90° are not included on this study.

1.8 Contributions of the Study

Our primary focus was on enhancing the performance of both the generator and discriminator network within our proposed FFRAD-GAN model.

- We incorporated channel attention mechanisms into the encoder, as well as self-attention mechanism into the decoder section of U-net based generator network.
- We employed a relative average based discriminator within the discriminator module to enhance its performance.
- We used L_2 loss mechanism along with multi-scale pixel wise loss and total-variation regularization loss.
- We present both quantitative and qualitative results using the Multi-PIE and CAS-PEAL-R1 datasets, demonstrating the frontalization of non-frontal faces with pose angles up to 60° for CAS-PEAL-R1 and 90° for Multi-PIE dataset.

1.9 Organizations of the Study

This study comprises seven chapters, which are basically highlighted below.

Chapter One: It provides an overview of the necessity of the study, the issues and a problem that motivated us to do this research.

Chapter Two: It describes literature reviews, deep learning and machine learning algorithms, image to image translation, face recognition and a related works done before on face frontaliazation tasks.

Chapter Three: It describes research methodologies including dataset collection, preprocessing of data, tools utilized for development, baseline work, and metrics for evaluation.

Chapter Four: It provides a detailed description about architecture of the proposed model, encompassing the design of the generator network and discriminator networks, the model learning functions, and the pseudocode outlining the training process.

Chapter Five: It goes into implementing the model, presenting a code samples illustrating the proposed solution, specifying the hyperparameters utilized, and describing the experiment class.

Chapter Six: It describes the proposed method outcomes and compares them to previous studies.

Chapter Seven: It offers a summary of the research findings and suggestions for future research.

CHAPTER TWO

2. LITERATURE REVIEW

2.1 Overview of Face Frontalization

Face frontalization is an approach of computer vision which generates a canonical-view of a face from an image in which the face is at various view or pose angle. It involves transforming a face image captured in any pose into its canonical frontal-view, which contains a better facial attribute (Yin et al., 2020; Luan et al., 2020,).

Frontal facial image is an image of a person's face that is captured from a straight-on angle, with face of the subjects directly facing the camera. This viewpoint guarantees the symmetry of the facial features and the correct positioning and visibility of all facial landmarks, such as the mouth, nose, and eyes. Because frontal facial images consistently and accurately depict a person's facial features, they are commonly utilized across various applications.

2.2 Face Frontalization Techniques

2.2.1 Traditional Face Frontalization Techniques

Traditionally research on frontalization of face has been done based on warping of 2D/3D local texture (Hassner et al., 2015;Asthana et al., 2011;Jeni & Cohn, 2016; X. Yin et al., 2017) and statistical modeling(Wenchao Zhang et al., 2005; Sagonas et al., 2015; Sagonas et al., 2017).

Among face frontalization by traditional approaches, (Jeni & Cohn, 2016) is one of approaches which proposed face frontalization based on 2D/3D local texture warping. Using a fast cascade regression technique, it first evaluated the visibility and position of a large number of markers. Next, it adapted a segmented 3D model to reconstruct the facial shape. Finally, it used the reconstructed 3D shape to compute a standardized views of the eye, helps in estimating 3D gaze.

(Sagonas et al., 2017) is used statistical modeling approach for frontalization of non-frontal face. It proposed a method for simultaneously reconstructing frontal views and localizing landmarks using only a limited set of frontal images, with the frontal facial image determined as having the lowest rank among various poses. They constructed a model that could recover both the frontalized face and facial landmarks concurrently.

2.2.2 Techniques for Face Frontalization using Deep Learning

Computer vision pertains to the extraction of information about a scene through the analysis of images depicting that scene (Rosenfeld, 1988). It's a domain within computer science that enables machines, like computers, to interpret and make decisions based on visual data. Just like our eyes and brains work together to understand the world around us, computer vision involves teaching machines to analyze and make sense of visual information. It encompasses techniques for obtaining, manipulating, interpreting, and comprehending digital images, along with extracting complex data from them. This field employs artificial intelligence, machine learning, and image processing techniques to make computers to recognize patterns, objects, and even behaviors in images or movies.

Machine learning (ML) is the scientific field that grants machines the capacity to improve their performance by learning without strict programming(Samuel, 1959). Instead of relying on rigid sets of rules, machine learning algorithms use statistical methods to allow computers to enhance their performance on a particular task as they gain experience over time. The rise of machine learning (ML) has been fueled by two key advancements: big data technology and high-performance computing. This powerful combination has opened doors to untangling complex data-driven processes in various fields. ML can now analyze massive datasets and uncover hidden patterns, leading to a deeper quantitative understanding of how these processes function within different operational environments (Liakos et al., 2018). Based on different learning capabilities and type of data used, machine learning techniques is classified as reinforcement learning, unsupervised learning, and supervised learning. Supervised algorithm is distinguished by its reliance on labeled datasets. In this approach, the input data is paired with corresponding output labels, essentially providing the algorithm with a pre-defined mapping between inputs and desired outputs. This labeled data serves as a training ground, allowing the supervised learning algorithm to learn the relation among inputs and outputs. This allows to make accurate prediction for unseen data (Liakos et al., 2018).

Unsupervised algorithms discover patterns and correlations within unlabeled data, identifying clusters or relationships within the dataset. Reinforcement learning algorithms acquire the ability to make decisions through interactions with an environment. It receives feedback through rewards or penalties, enabling it to learn the most effective actions to undertake in various circumstances. Machine learning has undoubtedly revolutionized computer vision, but challenges arise when dealing with complex data like images and videos. Traditional machine learning algorithms often struggle to extract meaningful features from this intricate data without significant guidance from domain experts. To address this limitation, deep learning is emerged. It is a powerful technique within machine learning, is particularly suited for handling complex data. Inspired by the human brain, it uses artificial neural networks to automatically uncover important patterns from massive datasets. This frees us from hand-picking these features ourselves and allows deep learning to achieve impressive results in computer vision tasks.

Deep learning is a branch of machine learning, gives computers the ability to learn and understand information in layers, much like the human brain. These layers, stacked on top of each other, allow the computer to process data in increasingly complex ways. It works similarly to how our brains make sense of the world. It can find hidden patterns in large amounts of data, just like the brain puts together information from different senses. Deep learning includes a wide array of methods, such as neural networks, hierarchical probabilistic models, and various algorithms for unsupervised and supervised feature learning (Voulodimos et al., 2018). It uncovers complex structures within vast datasets by employing the backpropagation algorithm to determine adjustments to a machine's internal parameters. These parameters are utilized to derive representations in every layer from the representations in the preceding layers (LeCun et al., 2015). Artificial neural networks (ANNs) are a type of deep learning inspired by the human brain. They're simplified models built like interconnected webs, mimicking how brain cells process information and learn. (Wu & Feng, 2018).

Deep learning provides a great improvement in computer vision, especially with convolutional neural networks. (LeCun et al., 1989), introduced a convolutional neural network for handwritten zip code recognition in 1989.CNN has gained the interest of both industry and academia in past few years due to its impressive achievements across different areas, like computer vision and natural language processing, among others (Z. Li et al., 2022). Based on CNN, in recent years, many effective generative models are emerged. Generative models learn the inherent patterns and structures within a given dataset and can subsequently produce new instances that resemble the original data. Among them GAN is well known generative model. Generative Adversarial Network (Goodfellow et al., 2014) is a generative model founded on a theory of a game between a generator and a discriminator neural network which attempts to find the Nash equilibrium between them and improves its ability to create more realistic samples.

Advancement of deep learning techniques like Convolutional Neural Networks (CNN) and Generative Adversarial Networks (GAN), increased achievement of face frontalization research (Huang et al., 2017; Kang et al., 2018; Al-Azzawi et al., 2018; Luan et al., 2020; Yin et al., 2020; Cen et al., 2022; Luo et al., 2022).

(Z. Zhang et al., 2019) introduced a Frontalization of Face with Convolutional Neural Network using Appearance Flow (A3F-CNN) which learns to construct a rich correlation between the side-view and frontal faces. After the correlation is established, frontal-view faces are generated through explicitly shifting pixels of non-frontal-view faces.

(Huang et al., 2017), introduced the Two-Pathway Generative Adversarial Network (TP-GAN) for producing photo-realistic frontal view images using deep learning approaches utilizing Generative Adversarial Network. This approach seeks to capture both overall structural features and finer local information concurrently using Convolutional Neural Network blocks and also four landmark-located patch networks to access local textures. This allows to generate frontal facial image with their improved feature representation.

2.3 Generative Adversarial Network (GAN)

Goodfellow first introduced Generative Adversarial Networks in 2014. The generative model examines a set of training samples and learns the likelihood distribution that produced them, allowing it to generate more realistic image examples that resembles to real images based on the estimated probability distribution. The GAN network comprises of two neural networks: the Generator G and the Discriminator D trained in a competitive manner, leading to a game-like scenario.

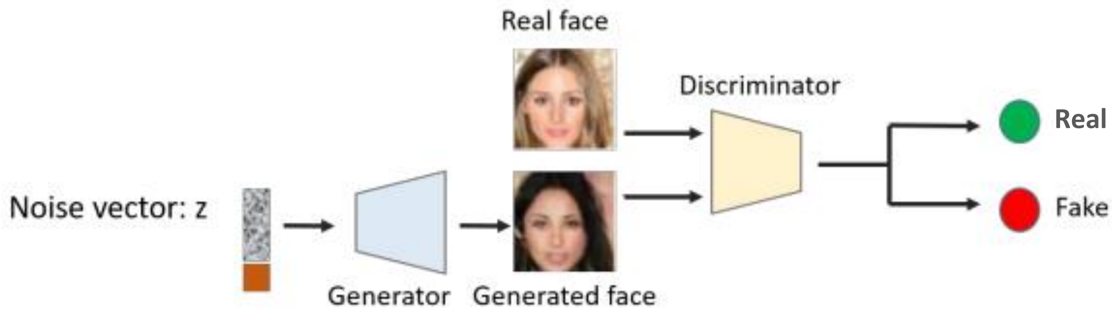

Figure 2. 1 Architecture of GAN

Source: adapted from (Kammoun et al., 2023)

The generator network generates new data, whereas the discriminator network determines whether a given piece of data is real or not. The training of the generator and discriminator occurs in an adversarial fashion. The generator aims to produce data that is so lifelike that it deceives the discriminator, while the discriminator strives to enhance its ability to identify fake data. This back-and-forth process forces both networks to improve, and eventually the generator becomes so good at generating data that it can deceive the discriminator even when the discriminator is very good at its job.

2.4 Generative Adversarial Networks Architectures

2.4.1 Conditional GAN

Generative models can learn the dataset's probability distribution and generate new samples from that distribution that is likely to the data they were trained in. However, there is no way to regulate the data being generated. In a conditional generative adversarial network proposed by (Mirza & Osindero, n.d.), the generator and discriminator networks used additional information, like class labels or other input features. This data is typically presented as either a one-hot encoded vector or a feature vector, which is then concatenated with the data's internal representation in a hidden space. The resulting vector is utilized as input to the generator. Similarly, the discriminator also receives the conditioning information as input, alongside the generated or real data sample. It leverages this information to discern between genuine and fake samples. This allows the cGAN to synthesize new data samples which are similar to the training data and have specific characteristics or properties. Many researchers used CGAN in research areas like translation of one image to another (Isola et al., 2017; Chai et al., 2018), text-to-image generation, face generation (Gauthier, 2015) and video generation (Wang et al., 2019).

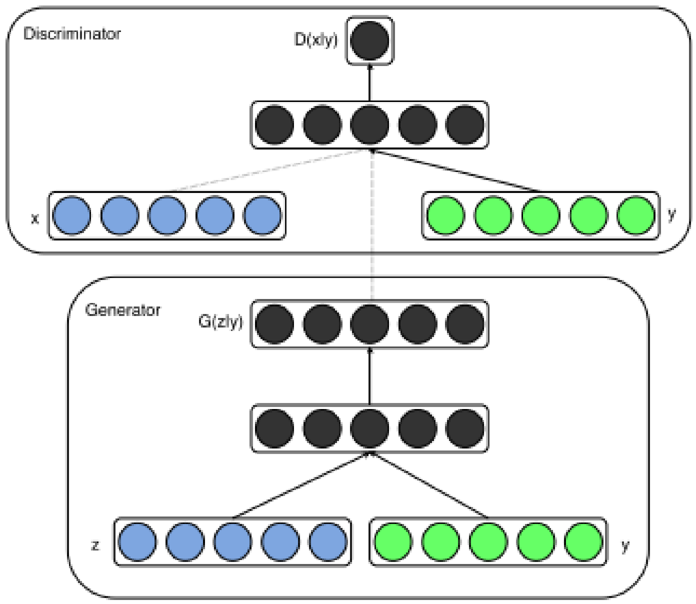

Figure 2. 2 CGAN architecture

Source adopted from (Mirza & Osindero, n.d.)

2.4.2 Least Squares GAN

While standard GANs treat the discriminator like a classifier, they use the sigmoid cross-entropy loss function. However, this choice of loss function could potentially result in vanishing gradients problem as the learning process progresses. (Mao et al., 2017) introduced the Least Squares Generative Adversarial Networks (LSGANs), in which the discriminator is trained with the least squares loss function to solve such problems. LSGANs can penalize the discriminator even when it correctly classifies a sample. When the generator is updated, the discriminator's parameters (decision boundary) are fixed and generate more gradients. Therefore, Penalization improves the generator with the capability to generate samples which are nearer to the decision boundary, thus resolving the vanishing gradient problem.

As proposed by (Mao et al., 2017) using the a-b coding scheme for the discriminator, the objective function for LSGAN is defined as follows:

$$\min_{D} V_{LSGAN}(D) = \frac{1}{2}\mathrm{E}_{\mathrm{X}} \sim p_{data}(x)\,[(D(x)-b)^2] + \frac{1}{2}\mathrm{E}_{\mathrm{z}} \sim p_z(z)\,[(D(G(z))-a)^2] \quad \textit{eq 2.1}$$

$$\min_{G} V_{LSGAN}(G) = \frac{1}{2}\mathrm{E}_{\mathrm{z}} \sim p_z(z)\,[(D(G(z))-c)^2] \quad \textit{eq 2.2}$$

Where:

- ✓ a is labels of fake data and b is labels of real data,
- ✓ c denotes the generator's (G) value endeavors to convince the discriminator (D) to accept as true for fake data.

2.4.3 CycleGAN

The objective of image-to-image translation is to learn the relationship between an input image and an output image using a training set of aligned image pairs. However, in many cases, such paired training data is not available. cycleGAN (Zhu et al., 2017) proposed a method for learning to convert an image from a source domain *X* to a target domain *Y* without the need for paired examples. It is a type of generative adversarial network for unpaired image-to-image translation. For two domains *X* and *Y*, CycleGAN (Zhu et al., 2017)learns a mapping $G: X \rightarrow Y$ and $F: Y \rightarrow X$. The innovation is in attempting to ensure that these mappings are reverses of each other and that both mappings function as bijections. This is accomplished through a cycle consistency loss that promotes $F(G(x)) \sim x$ and $G(F(y)) \sim y$. Combining this loss with the adversarial losses on *X* and *Y* yields the full objective for unpaired image-to-image translation.

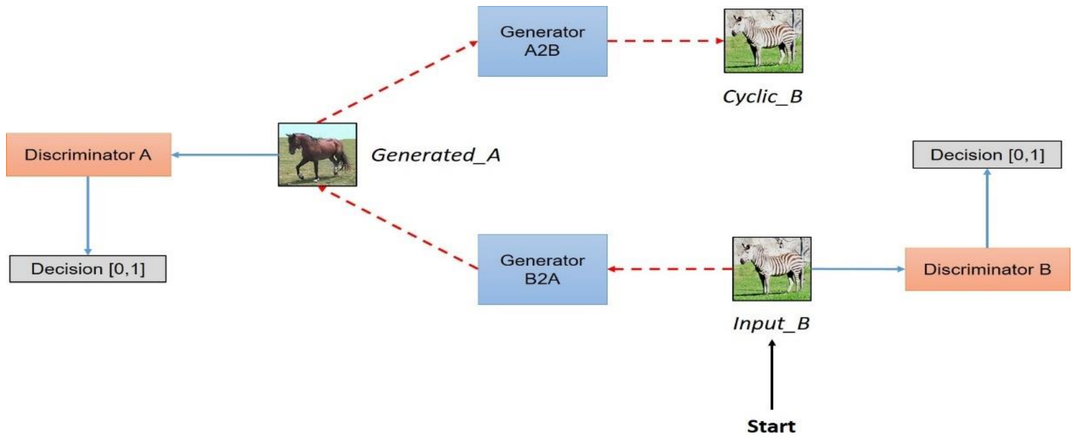

Figure 2. 3: cycleGAN Architecture

Source adopted from (Zhu et al., 2017)

2.4.4 Deep Convolutional GAN (DCGAN)

Deep Convolutional Generative Adversarial Networks (DCGANs)(Y. Zhang et al., 2018) represent a significant advancement in the field of generative adversarial networks (GANs), particularly in terms of stability and image quality. It leverages deep convolutional neural networks (CNNs) in both the generator and discriminator, moving away from the fully connected networks used in the original GAN architecture. It used transposed convolutional layers (deconvolutional layers) to upscale the input noise vector into a full-sized image in generator part and regular convolutional layers in discriminator part. DCGAN replaced pooling layers with strided convolutions in the discriminator and strided transposed convolutions in the generator to allow the networks to learn their own spatial down sampling or upscaling.

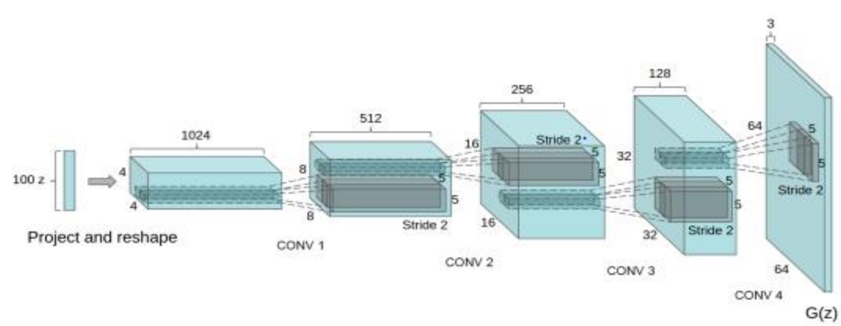

Figure 2. 4: DCGAN Architecture

Source adopted from (Y. Zhang et al., 2018)

2.4.5 Relativistic Discriminator

The relative discriminator is introduced by (Jolicoeur-Martineau, 2019) aiming to shift from true or false discrimination into a relative truth assessment. The discriminator calculates the likelihood that the provided real data is more probable to be realistic than randomly selected sample of fake data.

As proposed by (Jolicoeur-Martineau, 2019), The discriminator and generator loss functions of the Relativistic GAN can be expressed as:

$$L_D = -E_{(x_r, x_f)} \sim (P, Q)[log\ (sigmoid(C(x_r) - C(x_f)))] \qquad eq.\ 2.\ 3$$

$$L_G = -E_{(x_r, x_f)} \sim (P, Q)[log\ (sigmoid(C(x_f) - C(x_r)))] \qquad eq.\ 2.\ 4$$

Where x_r is the real data samples and x_f is fake data samples.

2.4.6 Relativistic Average Discriminator

The relative average discriminator, as introduced by (Jolicoeur-Martineau, 2019), assesses the input data by comparing it to the average critic score of samples of the opposite type. The discriminator calculates the average likelihood in which the provided real data is more probable to be realistic than synthesized data.

As proposed by (Jolicoeur-Martineau, 2019), The discriminator loss functions of the Relativistic average GAN can be defined as:

$$L_D = -E_{x_r} \sim P\left[log\left(\overline{D}(x_r)\right)\right] - E_{x_f} \sim Q\left[log\left(1 - \overline{D}(x_f)\right)\right] \qquad eq.\ 2.\ 5$$

$$\text{Where:} \quad \overline{D}(x) = \begin{cases} sigmoid\left(C(x) - E_{x_f} \sim Q\ C(x_f)\right) & \text{if } x \text{ is real} \\ sigmoid\left(C(x) - E_{x_r} \sim P\ C(x_R)\right) & \text{if } x \text{ is fake} \end{cases}$$

We employed the relative average based discriminator (Jolicoeur-Martineau, 2019) to assess the likelihood that a given frontal facial image is, on average, more authentic than a synthesized frontal image, leveraging prior knowledge heavily.

2.5 U-Net

U-Net is a deep learning architecture that was first proposed by (Ronneberger et al., 2015). It comprises two paths: contracting and expanding paths. The expansive path consists of decoder blocks that are responsible for decoding the encoded information. It leverages information from the contracting path through skip connections to generate a segmentation map. Conversely, the contracting path adopts a convolutional network architecture, including encoder blocks that capture contextual details and minimize the input's spatial resolution (Ronneberger et al., 2015).

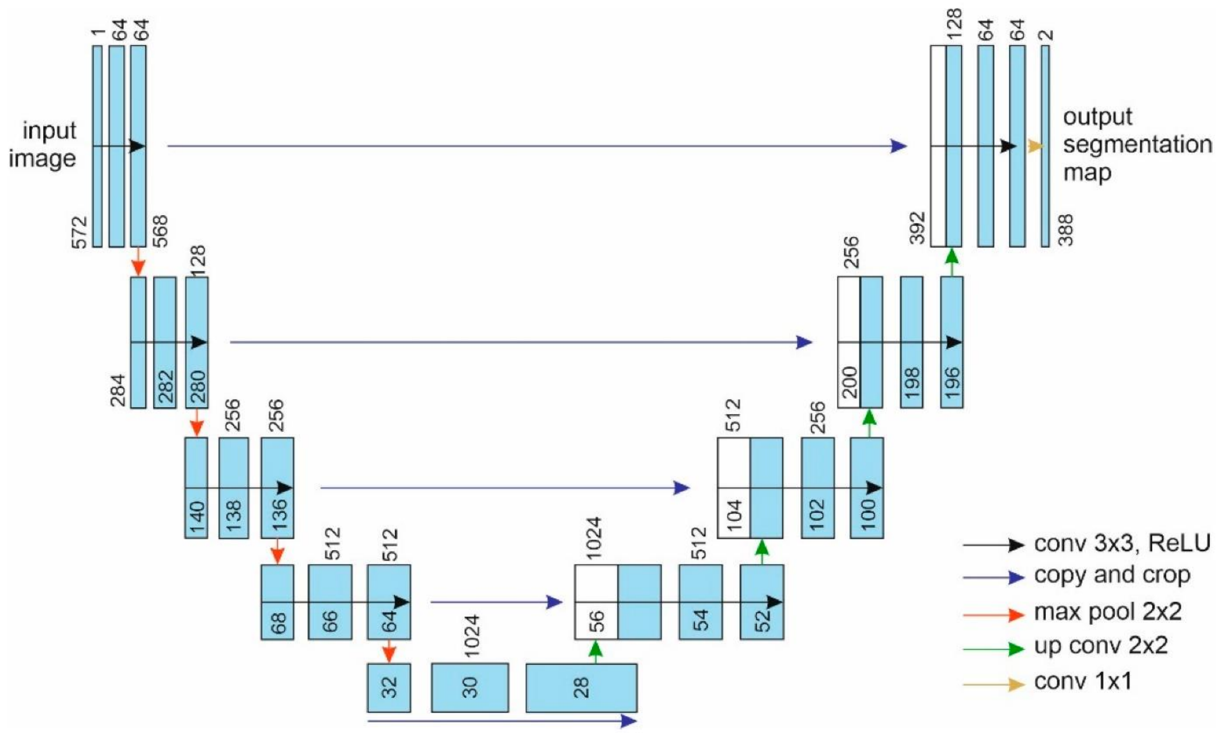

Figure 2. 5: U-net architecture

Source adopted from (Ronneberger et al., 2015)

2.6 Translation of Image to Image (I2I)

For instance, if we possess a selfie image and aim to convert it into a more artistic form, such as drawing by cartoon, this type of research is referred to as image-to-image translation. The aim of image-to-image translation (I2I) is to convert images from one category to another while preserving the content representations (Pang et al., 2022). Image-to-image translation methods can be implemented using either supervised training or unsupervised training settings. The supervised I2I translation requires a set of pair of samples datasets to convert source images to target images. Assembling a dataset with all matched samples can improve translation performance with supervised I2I translation, despite the difficulty and effort involved. Unsupervised image-to-image (I2I) translation tackles the challenge of translating images from one domain (source) to another (target) without relying on paired training examples (Shukla et al., 2019). I2I translation is gaining popularity right now because it could be utilized in a various applications of computer vision such as colorization of photos, frontalization of face and face recognition.

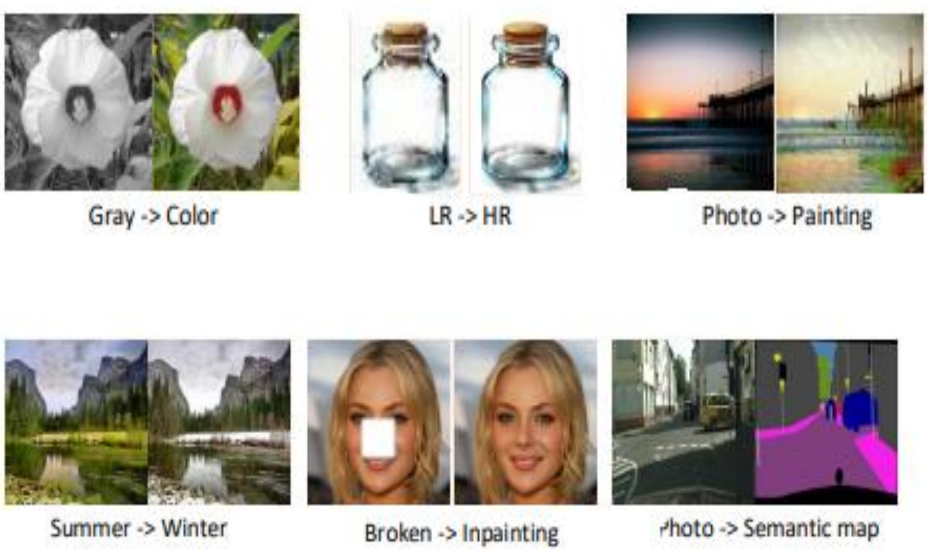

Figure 2. 6 Translation of image

Source adopted from (Pang et al., 2022)

2.7 Face Synthesis

Image generation has gotten a significant attention in recent years. Among applications in industry, one of them is synthesizing a facial image from a different perspective while preserving its identity. Face synthesis is the process of creating realistic and diverse images of human faces from scratch or from some input. Deep face synthesis models can be classified into three categories depending on the learned mappings between input and output images (Y. Shen et al., 2018). These are one-to-one, many-to-one, and many-to-many mapping. One-to-one mapping is a method of changing a face from one style to another, for example, from a sketch to an image, from low-resolution to high resolution, or from the visible to the infrared spectrum. Many-to-one mapping is a method of transferring a face into a single view(output) using a larger variation(multi-view) of input data, such as face frontalization which transforms multi-view face image to its frontal view (Y. Shen et al., 2018).

2.8 Face Recognition

The human face is a window into a person's internal state. It can convey a wealth of information, including a person's current mood, their intentions towards others, and their level of focus. Beyond these emotional cues, facial features also play a crucial role in personal identification. Facial recognition is a subfield of computer vision concerned with automatically identifying individuals from digital images or videos. This task requires the system to not only detect the presence of a face but also extract and analyze its unique characteristics. To achieve accurate recognition, the system must be robust against variations in factors like lighting, facial expressions, pose, aging, and even minor image transformations. But image translation, rotation, scaling, and pose, is a difficult task in face recognition (Parmar & Mehta, 2014).. It has emerged as one of the most extensively utilized biometric techniques for identity authentication, finding applications in sectors including military, finance, public security, and even in everyday life.

2.9 Facial Attribute Manipulation

The growing trend of sharing pictures and portrait photos on the internet drives the fast growth of face editing software. The ability to alter intrinsic facial features like age and gender as well as add or remove facial accessories like eyeglasses and facial hair makes manipulating facial attributes attractive (Chen et al., 2019). Because facial attribute manipulation offers so many opportunities for both research and practical application, it has attracted a significant interest. Early research (Thies et al., 2015; Yeh et al., 2016) focuses on particular characteristics of aging, beautification/de-beautification, expression changes, and facial hair generation. These methods are tailored to particular tasks and necessitate background information that isn't relevant to brand-new editing assignments. With advancement of deep neural networks, particularly the rise of generative adversarial networks, numerous frameworks for manipulating general face attributes have been proposed (Perarnau et al., n.d.,2016; W. Shen & Liu, 2017;Shu et al., 2017).These approaches treat the editing of facial attributes as an unpaired learning task, so they can handle multiple attributes by simply altering the data.

2.10 Related Works

Huang et al., 2017, proposed a Two-Pathway Generative Adversarial Network (TP-GAN) designed for generating photorealistic frontal view. This approach aims to capture global and local features. To address local textures, four patch networks centered around landmarks are utilized. They combined face domain knowledge with prior knowledge from data distribution to precisely retrieve the information that is lost when projecting a 3D object into a 2D image space. However, the synthesized image doesn't have finer details and becomes blurry especially with larger poses.

Luan et al., 2020, proposed a Geometry Structure Preserving based GAN (GSP-GAN), for frontalization and recognition of faces with multi-pose angles. The model's generator uses a conventional auto-encoder architecture, with the encoder extracting identifying information and the decoder generating the associated frontal facial image. They also utilized self-attention in the discriminator of model to maintain face's geometric structure. However, it lacks the capability to extract spatial context features within the generator network for image synthesis.

Yin et al., 2020, proposed Dual-Attention Generative Adversarial Network (DA-GAN) aims to achieve photorealistic frontalization of face by simultaneously acquiring contextual dependencies and local consistency. They presented a generator based on self-attention to combine local features with their distant dependence to produce improved feature representations. Face-attention based discriminator with four separate discriminators are utilized to highlight local aspects of face regions. However, it does not focus on the relevant facial attributes and does not recover well the relevant facial attributes like eyeglass, skin and hair in extreme poses above 45˚.

(Rong et al., 2020) This work proposes a novel Generative Adversarial Network (GAN) named FI-GAN specifically designed for face frontalization. FI-GAN focuses on improving the performance of facial recognition systems when dealing with faces captured from large poses. It achieves this by incorporating a unique step within the generation process: mapping intermediate features, extracted during the network's operation, into the frontal view. This allows FI-GAN to not only generate realistic frontal faces but also ensure these generated faces retain the critical features necessary for accurate face recognition, even with significant pose variations in the original image. They include a Feature-Mapping Block for assessing the difference between side-view and frontal facial features, and a feature discriminator for distinguishing between side-view and frontal face features. However, this work only focuses on addressing the variation of yaw angle and does not consider face images with large pitch angles.

Cen et al., 2022, this research introduces a novel Generative Adversarial Network (GAN) called PM-GAN for frontalizing faces. PM-GAN leverages a pre-trained feature fusion module (FFM) to address the challenge of limited feature diversity. This module strategically combines features extracted from the encoder with those obtained from a model pre-trained on a large dataset. This fusion process enhances the overall robustness and diversity of the features used for face frontalization, potentially leading to superior performance. However, it has a huge computing burden for the large image and increased computational cost of training.

(Cao et al., 2023) present a deep generative adversarial network utilizing multi-attention mechanism (DMA-GAN) in which they include a deep feature encoder in the generator leveraging attention mechanism and residual blocks. To hold global as well as local facial information they utilized four separate discriminators in the discriminator part of the network. However, this work only focuses on moderate poses (up to 45˚) and does not explore to more extreme poses above 45˚.

2.11 Summary of Related Works

Table 2. 1: Summary of related work

Related Paper	Methodology	Dataset used	Limitation
(Huang et al., 2017)	TP-GAN: Encoder-decoder network with landmark located patch networks.	Multi-PIE and LFW datasets	The synthesized image lacks fine details and becomes blurry especially with larger poses.
(Luan et al., 2020)	GSP-GAN: Auto-encoder as generator, Self-attention block-based discriminator	CMU and Multi-PIE datasets	Lacks capability to extract spatial context features in the generator network to synthesize image.
(Yin et al., 2020)	DA-GAN: Self-attention-based generator and Face-attention based discriminator	Multi-PIE and CAS-PEAL-R1 datasets	It does not focus on relevant facial attributes and does not recover well the relevant facial attributes like eyeglass, skin, and hair in extreme poses above 45˚.
(Rong et al., 2020)	FI-GAN: Feature-Mapping Block in generator, Feature discriminator in discriminator network	Multi-PIE and CFP datasets	It only focuses on addressing the variation of yaw angle and does not consider face images with large pitch angles.
(Cen et al., 2022)	PM-GAN: Feature fusion module (FFM) in generator, multi-domain identification in discriminator	Multi-PIE and CAS-PEAL-R1 datasets	It has a huge computing burden for the large image and increased computational cost of training.

(Cao et al., 2023)	DMA-GAN: Attention mechanism and residual blocks in generator, Four separate discriminators in the discriminator	CAS-PEAL-R1 dataset	It only focuses on moderate poses (up to 45˚) and does not explore to more extreme poses above 45 ˚.

Finally, some studies (Huang et al., 2017; Yin et al., 2020)do not adequately focus on the most important facial features that effectively describe the subject. As a result, the frontalized facial images frequently lack fine details and important facial features, and they tend to become blur. To solve this issue, our FFRAD-GAN incorporates a hybrid attention and a relative average based discriminator with in a GAN. This method seeks to improve focusing on the most essential facial features while reducing blurriness in frontalized facial images.

Another works (Rong et al., 2020; Cao et al., 2023) focuses only on addressing the variation of yaw angles and pose angle up to 45˚. But in real world, the acquired facial image pose angle are being increased to extreme poses greater than 45˚ and the pitch angles have a great impact on quality of frontalized facial image. This paper proposed to solve this problem by considering pitch angles in addition to yaw pose angles while frontalizing faces and extreme pose angles up to 90˚.

CHAPTER THREE

3. RESEARCH METHODOLOGY

3.1 Overview of Chapter

To realize the research goals outlined earlier, this chapter delves into the specific methodologies, procedures, and techniques employed. This includes a detailed exploration of the datasets, augmentation of the data, the data preprocessing steps, an overview of the baseline methods, and the development tools. Finally, the evaluation metrics chosen to assess the model's performance.

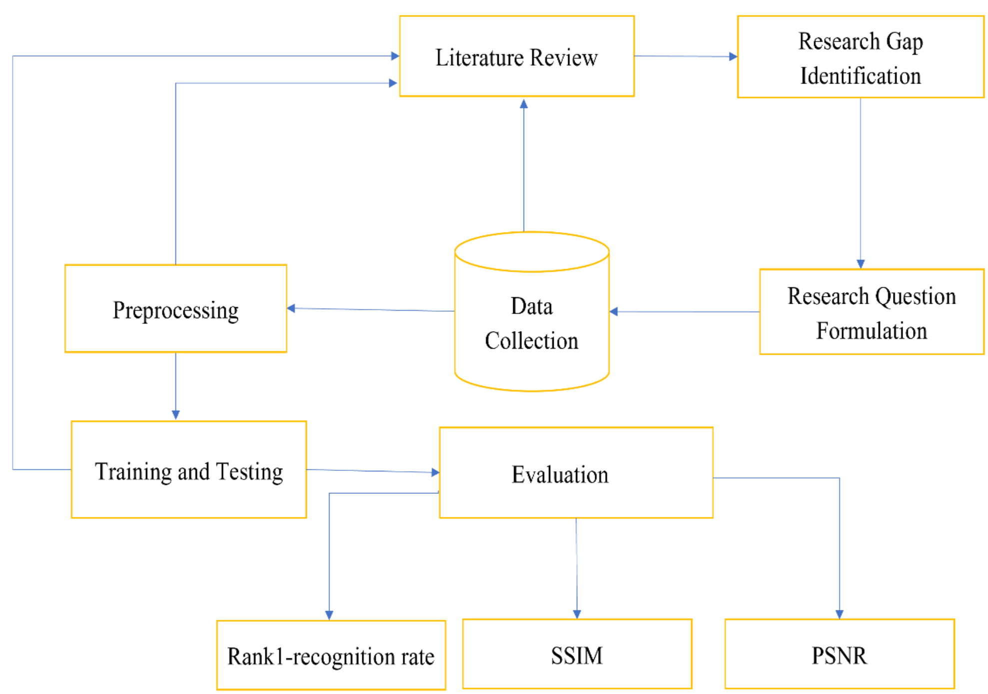

Figure 3. 1 Overall process

3.2 Datasets

This research addressed the challenge of face frontalization by leveraging the power of deep learning techniques. For learning, deep learning models mainly depend on data. Thus, data plays a crucial role in research. In order to carry out face frontalization using Generative Adversarial Network utilizing combination of attention and relative average based discriminator, we were used Multi-PIE(P. Li et al., n.d.-a), and CAS-PEAL-R1 (Wen Gao et al., 2008) datasets for training and testing. We selected those datasets to use on this study due to they contain a variety of facial image with different pose, illumination, expression, gender and age variation.

3.2.1 CAS-PEAL-R1 Dataset

The CAS-PEAL-R1 dataset (Wen Gao et al., 2008) is a comprehensive collection of Chinese faces compiled by ICT-ISVISION Joint Research and Development Laboratory (IDL) with controlled pose, expression, accessory, and lighting conditions. Each individual has 18 images with looking left, right, up and down direction. It is a rich resource containing 30,863 grayscale images of 1,040 individuals (595 men and 445 women) with various age level. Among images on CAS-PEAL-R1 dataset, 16,844 facial images are used for training and 4195 facial images are used for testing.

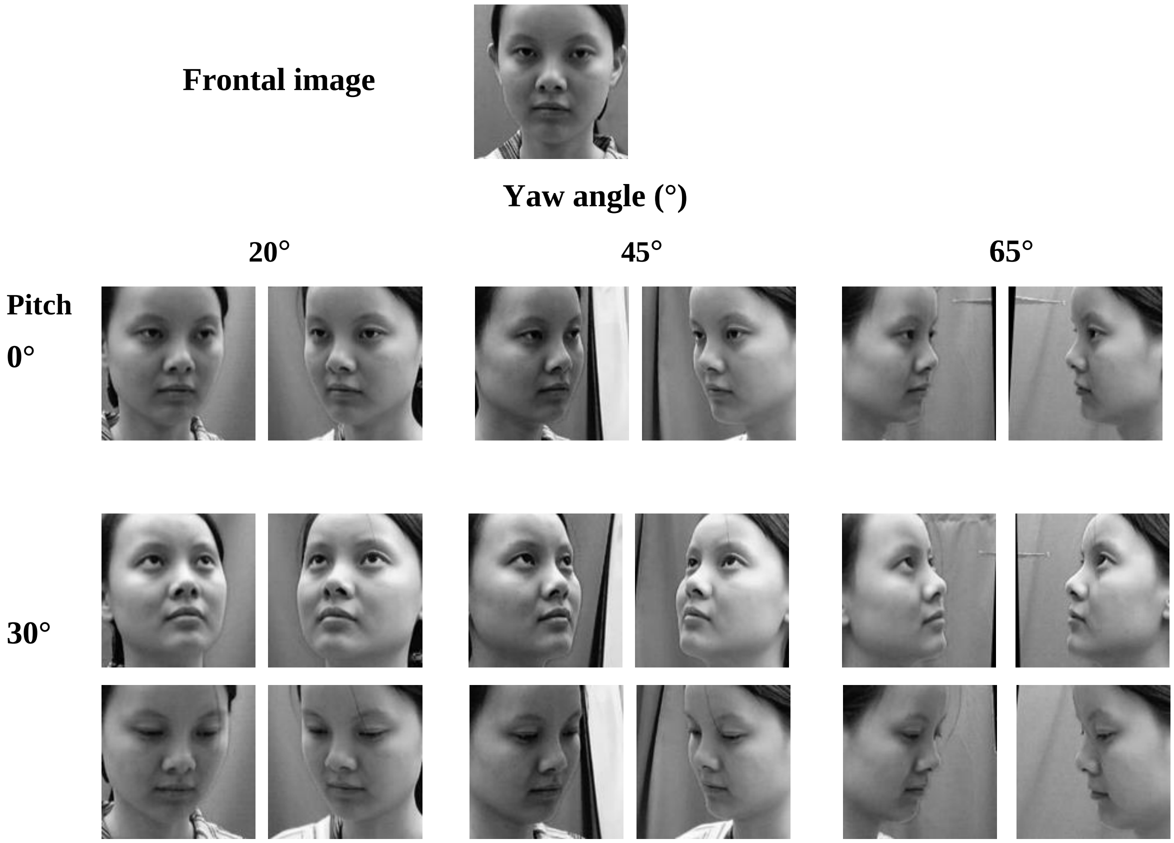

Figure 3. 2 CAS- PEAL- R1 (Wen Gao et al., 2008) dataset samples

3.2.2 Multi-PIE Dataset

The Multi-PIE dataset (P. Li et al., n.d.-a) is a dataset for facial pose and recognition research provided by university of Chinese academy of science. It includes images of 229 individuals with various yaw, pitch, attribute, illumination, and accessory. It provides images of subject ranging from -90° to 90° pose angles with different age and gender. Among the images in Multi-PIE dataset, we used 20,000 facial images selected randomly for training and 3000 images for testing.

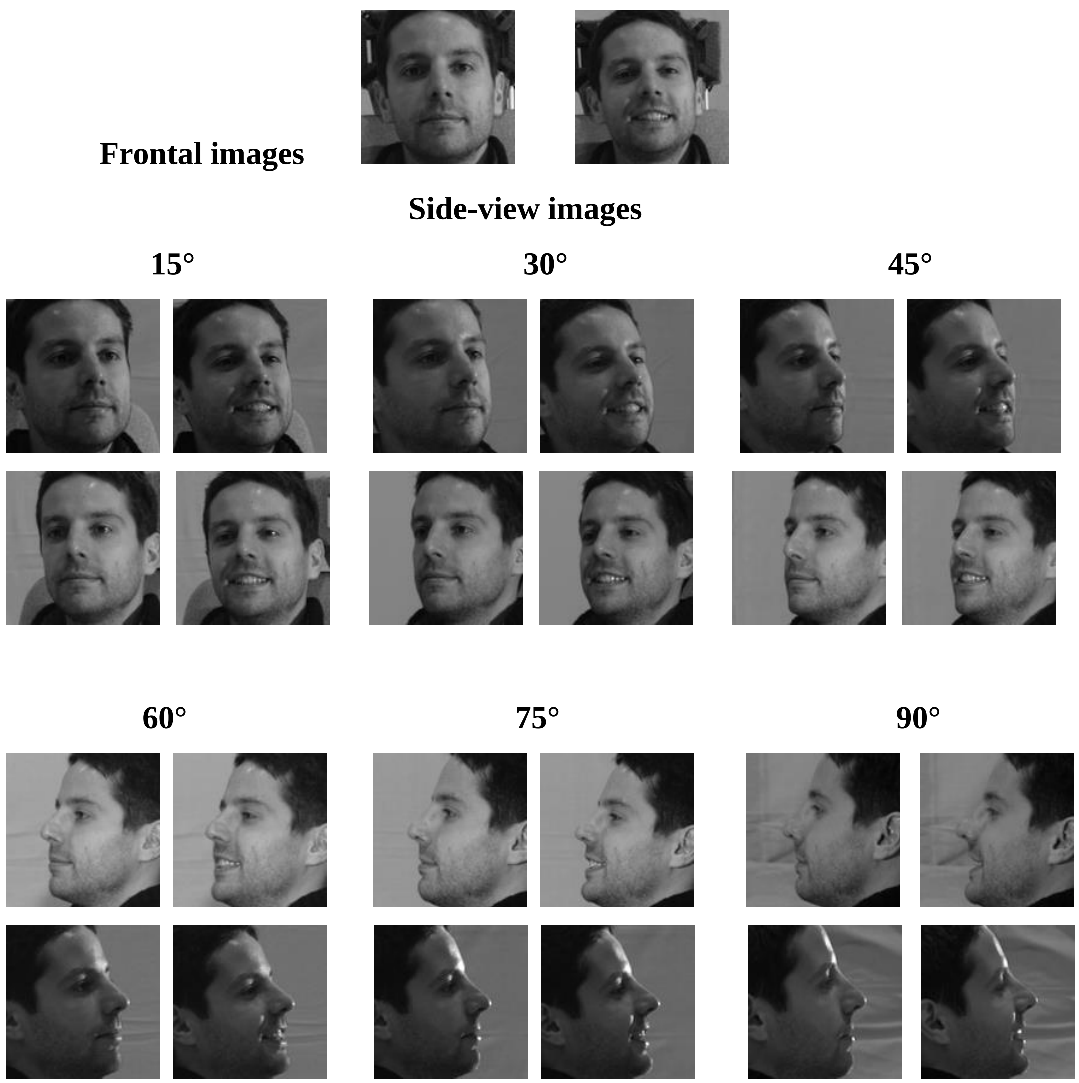

Figure 3. 3 Multi-PIE (P. Li et al., n.d.-a) dataset samples

3.3 Pre-processing Techniques

3.3.1 Data Augmentation

A training dataset's size and variety can be artificially increased using data augmentation techniques. The augmented dataset can be used to train model to make it more resilient and adaptable to different scenarios. This helps in avoiding overfitting. We augmented the training data using a set of techniques to increase its variability. This enrichment process aimed to improve the model's generalization capabilities and achieve better performance. In our study we applied the following data augmentation strategies to increase our dataset by three times.

- Brightness and Contrast adjustment: We altered the brightness and contrast of our image dataset by brightness between -50 and 50, contrast between 0.8 and 1.2.
- Scale: We also scale our images with scaling factor range between 0.8 and 1.2
- Gaussian blur: We applied gaussian blur by kernel size of 5x5 to simulate out of focus effects.

3.3.2 Normalization

Normalization is a method of putting data into a consistent and comparable format. This usually entails scaling the data to a given range or distributing it based on a particular statistical feature. This enhances model performance, optimizes algorithms and reduce training time. In this study, the dataset is normalized within the range of -1 to 1, which implies that the highest and lowest values for each attribute are constrained to -1 and 1, respectively.

$$x = \frac{y-\mu}{\delta} \qquad \textit{eq 3.1}$$

Where, y represents the input image within range of 0 to 1, while μ denote the mean, and δ denote the variance. In this context, the mean is set to 0.5 and the variance to 0.5.

3.3.3 Resize

Image resizing is a crucial pre-processing step. It involves adjusting the dimensions of the input images to match the model's expected input size. Since deep learning requires a large number of instances, the model performs additional computations to handle larger images. Resizing to a smaller size can minimize the computational resources required and greatly speed up training. We resized the images in to $128x128$ to minimize the complexity of computation.

3.4 Development Tools

To design and execute this study, a variety of software and hardware tools were employed.

3.4.1 Hardware Tools

The following hardware tools were utilized for implementing the research.

Table 3. 1: Hardware tools

No.	**Tools**	**Purpose**
1	RAM	To accelerate the training process and improve the performance
2	Hard Disk	For storing data and models
3	GPU	To enhance computation and speed up the training process,

3.4.2 Software Tools

Software tools help with various aspects of the research process, such as preprocessing of data and model development, as well as visualization and deployment.

Anaconda[1]***:*** It is a popular open-source installation of programming languages for scientifical computation, data analysis, and machine learning. It comes with a large number of pre-installed packages and libraries, including prominent ones like NumPy, pandas, SciPy, Matplotlib, and scikit-learn, among others.

Jupyter notebook[2]***:*** It's an open-source web application that enables users to create and share documents containing live code, equations, visualizations, and narrative text.

[1] https://anaconda.org/

[2] https://jupyter.org/

TensorFlow[3]***:*** It is an open-source machine learning framework which provides a comprehensive set of tools, libraries, and community resources to help researchers and developers build and deploy machine learning models.

PyTorch[4]: It is a powerful and versatile framework for deep learning research and development, offering flexibility, ease of use, and strong community support.

3.5 Baseline Works

In order to see how well our proposed model works, we compared it to other more recent models that use generative adversarial networks (GANs) for frontalization of face. DMA-GAN (Cao et al., 2023) was selected as a baseline of this study based on generated results without requiring the prior facial knowledge like type of pose when compared to others.

The DMA-GAN (Cao et al., 2023) model comprises of a single generator and four discriminators. The generator is built with a deep feature encoder with attention mechanisms and residual blocks. The discriminators are discriminators for the whole face, eye, nose, and mouth. The author of the baseline work emphasizes that combining the residual block and the attention mechanism could enhance the depth of layers of the network and finer facial features extraction. To hold both global and local facial information they utilized four separate discriminators in the discriminator part of the network.

3.6 Feature Extraction Network

We used ResNet18(He et al., 2015) for extracting features in order to calculate Rank1-recognition rate of our model. It is a convolutional neural network architecture that is widely utilized in applications of computer vision. It is noted for its deep structure with residual connections, which helps in training deeper networks more effectively (He et al., 2015). In face frontalization, the main aim is to frontalize the side view facial image of a subject into the high-quality frontal facial view which correctly identify the respected subject. We used cosine-distance metric to compute the similarity between the features of frontal facial image frontalized

[3] https://www.tensorflow.org/

[4] https://pytorch.org/

by our proposed model and the features of ground truth frontal facial image to calculate rank1-recognition rate.

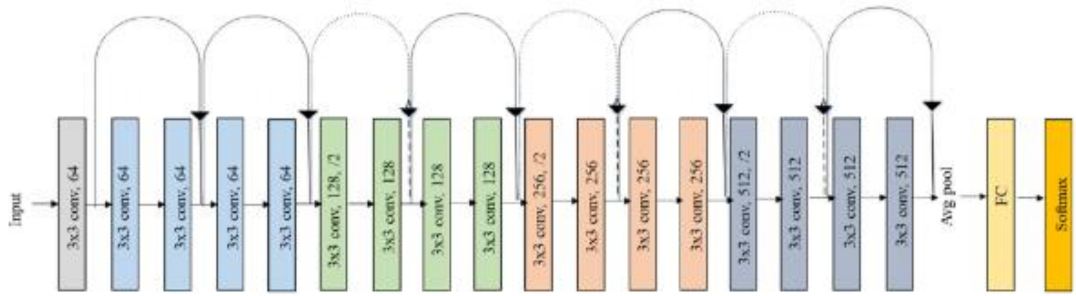

Figure 3. 4 ResNet18 architecture

Source adopted from(Ramzan et al., 2020)

3.7 Evaluation Metrics

In this study, we used objective metrics (rank-1 recognition rate, PSNR and SSIM) to assess the performance of the frontalization results.

3.7.1 Rank-1 Recognition Rate

The Rank-1 recognition rate is a popular metric for assessing the performance of image recognition algorithms. It indicates how frequently the algorithm assigns the highest probability to an image's true class.

It can be calculated as:

$$\boldsymbol{Rank_1 RR = \frac{R_1}{N}} \qquad \textit{eq. 3.2}$$

Where *N* represents the total number of identifications attempts, R_1 denotes the number of times the correct match is ranked first.

3.7.2 PSNR (Peak Signal to Noise Ratio)

This value indicates how closely a reconstructed image or signal matches the original, considering factors like detail preservation and noise levels. It is expressed as the ratio of a signal's maximum possible power to the power of disruptive noise affecting the quality of its representation. It is expressed in decibels (dB) and a higher PSNR indicates a good reconstruction which ensure a high image enhancement (Isa et al., 2015).

$$PSNR = 20log_{\,10}\left(\frac{MAX\_I}{\sqrt{MSE}}\right) \qquad eq.\ 3.\ 3$$

Where:

- ✓ MAX_I denote the maximum possible intensity value of a pixel in the image
- ✓ MSE denote the mean squared error between the original and the synthesized image. It is calculated as:

$$MSE = \frac{1}{M*N}\sum_{i=1}^{i=M}\sum_{j=1}^{j=N}\ (f(ij) - g(ij))^2 \qquad eq.\ 3.\ 4$$

Where *f* is the original, *g* is synthesized images and *M*N* is size of images.

3.7.3 SSIM (Structural Similarity Index Measure)

Structural Similarity Index Measure (SSIM) is a metric used to measure the similarity between two given images and quantifies image quality degradation caused processing of image. It compares original image and output image by extracting three key features: luminance, contrast, and structure.

$$SSIM(i,j) = \frac{(2\mu_i\ \mu_j + c1)(2\ \sigma_{ij} + c2)}{(\mu_i^2 + \mu_j^2 + c1)(\sigma_i^2 + \sigma_j^2 + c2)} \qquad eq.\ 3.\ 5$$

Where *i* and *j* are the two images being compared, μ_i, and μ_j are means of *i* and *j* respectively, σ_i^2 and σ_j^2 are the variance of *i* and *j*, $\sigma_{ij}$is the covariance of *i* and *j*, and c_1 and c_2 are constants to stabilize the division with weak denominator.

CHAPTER FOUR

4. PROPOSED ARCHITECTURE

4.1 Overview of Chapter

This chapter outlines an architecture of proposed method for face frontalization, employing a combination of attention and a relative average based discriminator (FFRAD-GAN). It includes graphical illustrations and mathematical formulations. The chapter is structured into four segments: the initial section provides an overview of the model's architecture, followed by an explanation of the generator and discriminator design in the second section. The third segment describes about learning functions of the model, and the concluding part explains the training pseudocode.

4.2 Model Architecture

The proposed approach named "FFRAD-GAN" is relies on the principle of Generative Adversarial Network (Goodfellow et al., 2020). The GAN network comprises of two neural networks: Generator G and Discriminator D train in a competitive manner. Our proposed approach used U-Net (Ronneberger et al., 2015b) in generator which consist of encoder-decoder structure with skip connection.

A self-attention module (H. Zhang et al., 2018) was added in the decoder part of generator to capture long-range contextual information. In encoder part of generator, we were used channel attention module (Hu et al., 2018) in the generator network to prioritize the most important facial features which essential for frontalization. We utilized a channel attention layer before a skip connection to selectively focus on important information before it is propagated to the decoder part of the generator. In order to deceive the discriminator, the generator synthesizes an artificial image by using information from the real image.

In discriminator network we used relative average discriminator (Yuan et al., 2020) to enhance the performance of the discriminator. This method goes beyond simply classifying something as true or false. Instead, it judges how real something is compared to something else known to be fake. It is the discriminator's responsibility to differentiate between images created by the generator (fake images) and a real image. The discriminator undergoes continuous adjustments to evaluate the images generated by the generator, while the generator continuously learns to produce images that are indistinguishable from genuine ones.

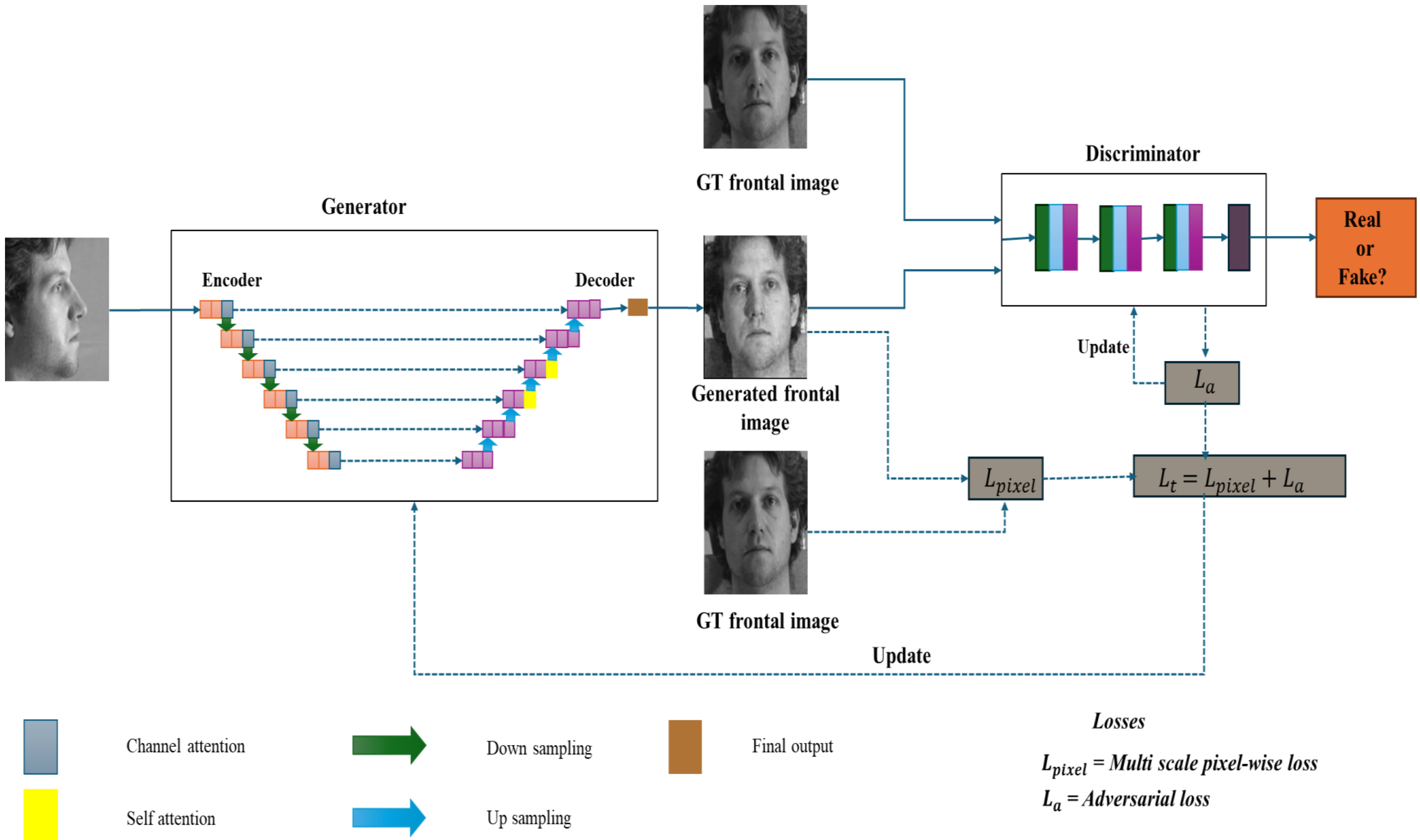

Figure 4. 1 Overall architecture of the proposed model

In our proposed FFRAD-GAN approach, we used channel attention in encoder block and self-attention in decoder block of U-net based generator in order to improve its capability to synthesize a realistic frontal facial image to fool the discriminator.

In the GAN framework, self-attention allows the generator and the discriminator to effectively capture long-range relationships within image features(H. Zhang et al., 2018). It is essential in face frontalization tasks to maintain the relationships between important facial features like eye, nose and mouth. It also preserves fine details of face like skin texture, shape of lips and sharpness of the eye which allows to produce more realistic frontalized facial image. Self-attention also reduces frequent GAN distortions and blurriness by offering a mechanism for refining and improving generated features depending on their global context. This allows to produce a cleaner and sharper frontalized images. To leverage these significant features, we used self-attention in decoder part of the generator.

Channel attention is a method employed in deep convolutional neural networks (CNNs) to concentrate on the most relevant channels within a feature map. This enhances network performance by reducing the processing of noise and irrelevant information (Hu et al., 2018). It allows the model to prioritize on more important features of face like shape and texture of eye, mouth and nose, resulting in better preservation of these features in the frontalized facial image. To leverage these significant features, we utilized channel attention in encoder part of generator. Using channel attention in encoder part allows to build a strong foundation for the next layers of the model due to the encoder refines key features by focusing on the most important features of the face, which are crucial for accurate frontalization.

4.3 Generator Architecture

In GAN (Generative Adversarial Network), the generator network is depicted as the component responsible for producing synthetic data that closely mimics real data. It is typically structured as a neural network, often using convolutional layers, which takes random noise as input and outputs data samples that are intended to be indistinguishable from real data. Our generator utilized U-Net architecture (Ronneberger et al., 2015b) which consist of encoder-decoder structure with skip connection and hybrid attention (i.e., self-attention and channel-attention).

U-Net is an approach on deep learning first presented by (Ronneberger et al., 2015), comprises two main pathways: contraction and expansion. The expansion pathway involves decoder layers tasked with decoding encoded information and integrating details from the contraction pathway via skip connections to produce a segmentation map. Conversely, the contraction pathway follows a convolutional network structure, featuring encoder blocks that capture contextual information and minimize the input's spatial dimension (Ronneberger et al., 2015).

The encoder part of our generator contains Convolution, Batch normalization as well as ReLU, the decoder part of generator contains Transposed convolution, Batch normalization and ReLU.

In order to learn more condensed and conceptual representations of the image, the bottleneck layer in the generator network compresses the input and lowers the dimensionality of the feature maps. The generator part of the network also contains skip connections to retain high-resolution features from earlier layers, which is crucial for maintaining detailed facial information during the transformation process.

The self-attention block is utilized at the third and fourth layers of the decoder to enable the generator to acquire both local and global dependencies in intermediate feature maps. This can help in improving the spatial details and fine structures of the synthesized frontal faces. In all encoder layers, channel attention is utilized to allow the network to concentrate on most relevant features.

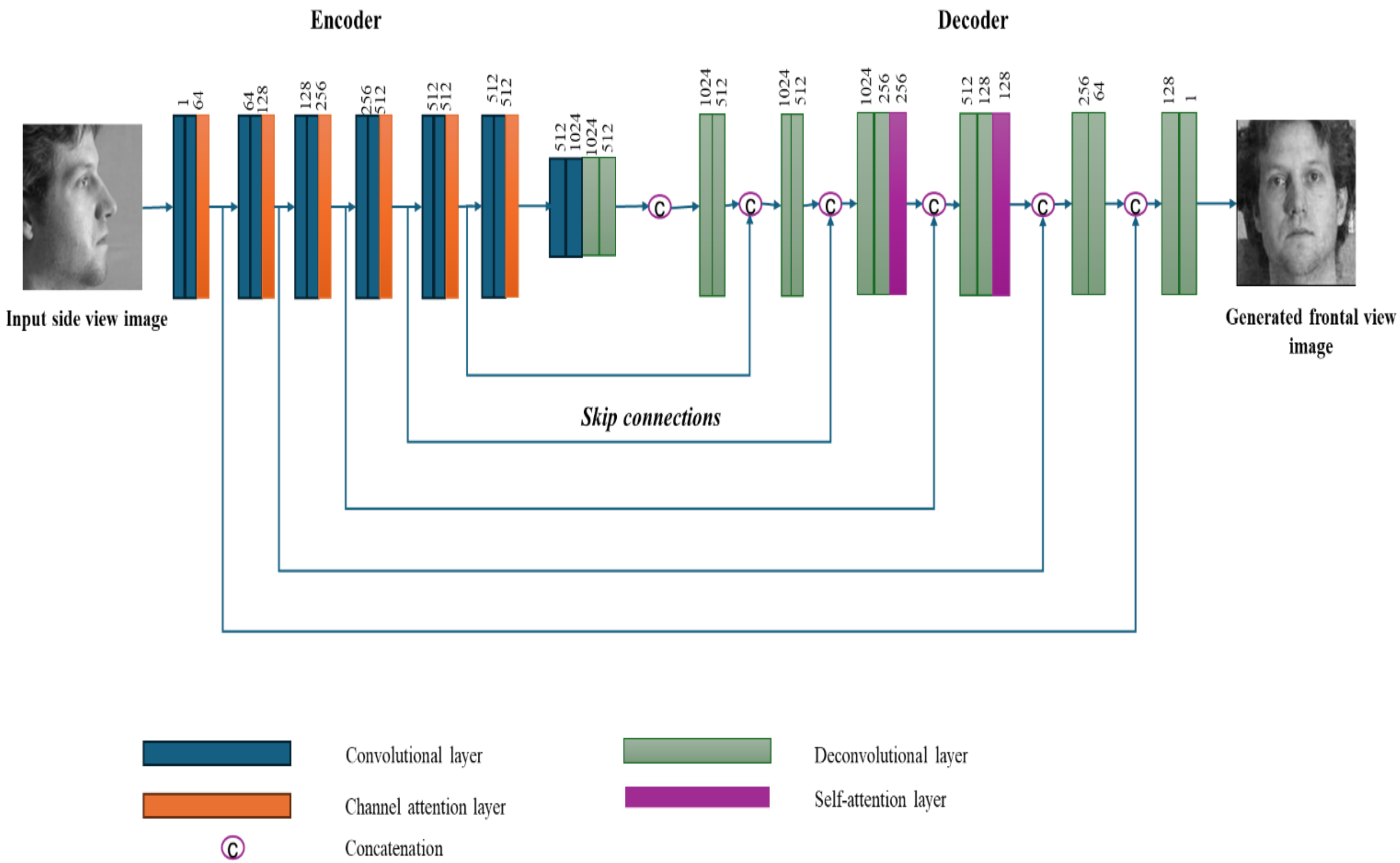

Figure 4. 2 Generator architecture

The encoder block in our proposed architecture, is a down-sampling block which contains six convolutional layers (conv1 – conv6) and channel attention block after each convolutional layer. Each convolutional layer has a stride of 2 and kernel size of 4. Batch normalization and ReLU activation function is used after each convolutional layer.

The decoder block is an up-sampling block which contains six deconvolutional layers (deconv1 – deconv6). Self-attention block is added on the third and the fourth layers. Each deconvolutional layer has a stride of 2 and kernel size of 4. ReLU activation function and Batch normalization are utilized after each deconvolutional layer except the deconv6 layer. Tanh activation function is used in the deconv6 layer.

4.3.1 Self-attention

Self-attention is a technique that originated to be used in machine learning, especially for tasks like natural language processing (NLP). It helps models understand the relationships between parts of a sequence and understand how different parts of the data relate to each other. Subsequently, this method found application in diverse fields like computer vision. The majority of image generation models relies on Generative Adversarial Networks (GANs) are constructed using convolutional layers. Due to convolution operates on information within a limited neighborhood, relying solely on convolutional layers to capture distant relationships in images is inefficient in terms of computational resources. (H. Zhang et al., 2018) proposed SAGAN which introduced self-attention in GAN framework in order to enable generator and discriminator to efficiently capture long-range relationships. SAGAN employs spectral normalization to reduce computational costs during training.

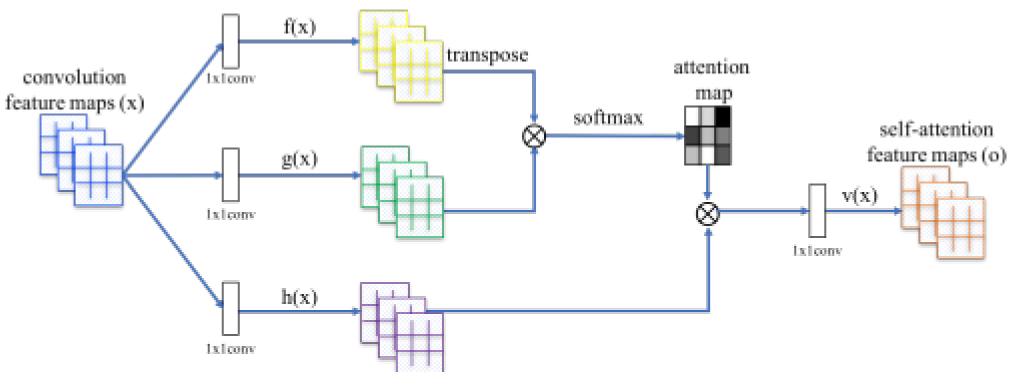

Figure 4. 3 Self-attention based GAN block.

Source adopted from (H. Zhang et al., 2018)

Self-attention enables the generator to acquire both local and global dependencies in intermediate feature maps. This can help in improving the spatial details and fine structures of the synthesized images.

4.3.2 Channel Attention

The channel attention is first introduced by SENet (Hu et al., 2018) to learn channel weights for each feature map. It is a method used in deep convolutional neural networks (CNNs) to focus on the most relevant channels in a feature map. This can help to enhance the network performance by reducing the amount of noise and irrelevant information that is processed. A channel attention map is produced by leveraging feature inter-channel relationships to concentrate on meaningful aspects with in an input image as treating each channel of a feature map as a detector of features. The spatial dimension of the input feature map is compressed to efficiently compute channel attention, with average pooling used to aggregate spatial data (Woo et al., 2018).

Channel attention blocks are intended to selectively emphasize informative feature channels while suppressing less important ones. They accomplish this by calculating channel-wise attention weights, which are then applied to the feature maps. This method allows the network to concentrate on the most discriminating features for the task.

A method for calculating attention weights is usually the component of a channel attention block. This method frequently uses global average pooling and is succeeded by fully connected or convolutional layers to capture inter-channel dependencies. The original feature maps are then modulated using these attention weights, which successfully highlights significant features and suppresses noise or less informative channels.

The provided equation defines a function channel_based_statistic that computes a channel-based statistic for a given input tensor. The input tensor is expected to be a 3D numpy array with dimensions (K, H, W), where:
K is the number of feature maps; H is the height of the feature maps and W is the width of the feature maps

$$Z_c = H_{GP}(y_c) = \frac{1}{HxW} \sum_{i=1}^{H} \sum_{j=1}^{W} y_c(i,j) \qquad \textit{eq. 4.1}$$

Where $y_c(i,j)$ represent value on (i, j) and H_{GP} (.) represents global average pooling.

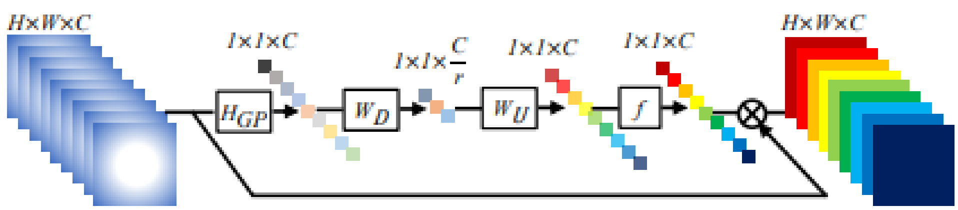

Figure 4. 4 Channel attention block

Source adopted from (Y. Zhang et al., 2018)

4.4 Discriminator Network

The discriminator network plays a crucial role in a GAN (Generative Adversarial Network) as it is tasked with distinguishing between real and synthetic data samples. It is typically implemented as a neural network, often using convolutional layers, which takes input data samples (both real and generated) and outputs a probability score indicating the likelihood that the input sample is real.

Our discriminator is relative average discriminator based on the work (Jolicoeur-Martineau, 2019) to allow our model to generate a better-quality frontal view facial image, and to enhance its performance. The generator is a network designed to produce a synthesized facial image, which is compared with the ground truth facial image by a relative average discriminator. This discriminator provides gradient information to the generator, enabling it to generate higher-quality images.

A relative average discriminator (RAD) is one form of discriminator used in Generative Adversarial Networks (GANs). Unlike standard GAN discriminators, which evaluate the absolute likelihood of an input being real, RADs focus on relative realism in comparison to real data.

Because relative average discriminator is so good at assessing the relative realism within a batch, we choose to utilize it instead of standard discriminator. It can effectively push truly frontal faces higher in the ranking compared to those with minor pose variations, even if all are somewhat frontal according to a standard discriminator.

4.4.1 Rationality for using Relative Average Discriminator

The relative average discriminator as introduced by (Jolicoeur-Martineau, 2019) evaluates the input data's critic by contrasting it with the average critic of samples of the opposite type. Based on (Jolicoeur-Martineau, 2019), the relative average discriminator enhances the performance of GAN 's discrimination and stability.

In a standard discriminator, the discriminator network produces the likelihood that a given sample is real, which can result to the discriminator becoming too powerful and providing poor gradient information to the generator; however, a relative average discriminator compares the realism of generated samples to real samples, with the goal of improving relative quality rather than absolute. This strategy provides more useful feedback to the generator.

A relative average discriminator also improves a gradient information to be provided to the generator. In standard GANs, the discriminator's gradients can become uninformative if it becomes too adept at distinguishing real from fake. This reduces the generator's ability to improve. While, the relativistic discriminator provides more richer gradients by assessing the relative realism. This allows the generator to acquire more information and meaningful gradients during the training process, resulting in a better learning process (Jolicoeur-Martineau, 2019).

Using relative average discriminator also encourage diversity of samples which reduces a model collapse. A standard GANs mainly tend to suffer by mode collapse, which occurs when the generator creates samples with little diversity in order to deceive the discriminator. However, a relativistic discriminator encourages a generator to produce a wider variety of samples in order to enhance its relative ranking, encouraging diversity and minimizing the chance of mode collapse.

Generally, relative average discriminator aligns with human perception as humans evaluate the quality of images in a comparative manner rather than an absolute one. The relativistic approach aligns better with this perceptual process. Therefore, the relativistic discriminator's comparative assessment helps produce samples that are more visually realistic and perceptually pleasing, as it better mimics human judgment of image quality.

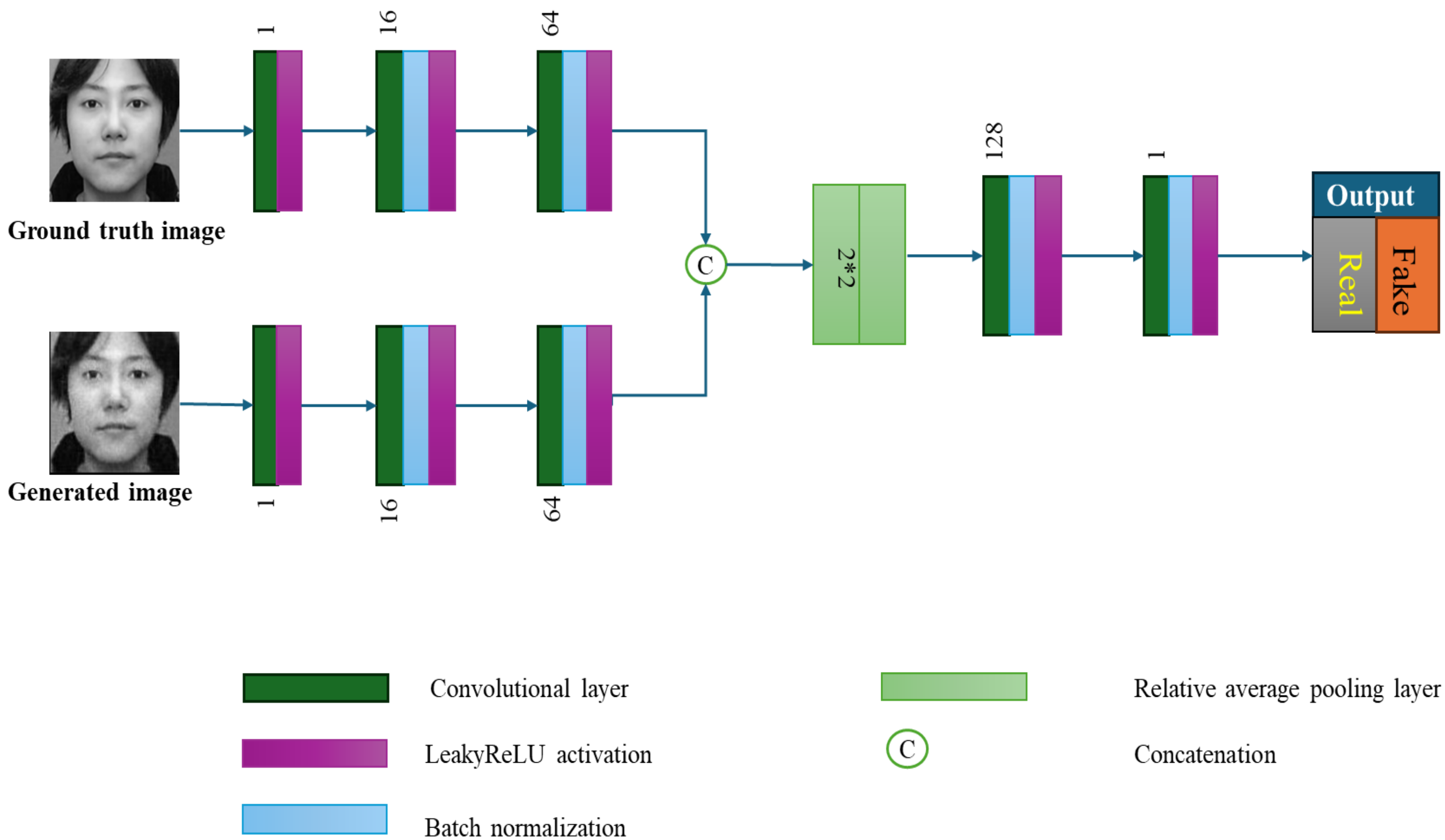

Figure 4. 5 Architecture of discriminator

4.5 Learning Functions of the Model

Planning a suitable learning function is thought to be quite important for producing a better result. Adversarial loss, multi-scale pixel wise loss, l_2 loss, and total variation regularization losses are used in this work.

4.5.1 L_2 Loss

In GAN-based face frontalization, the generator network attempts to transform a profile image into a realistic-looking frontal view. The landmark-based loss function improves this process by comparing the predicted facial landmarks in the generated frontal face with the corresponding landmarks in the real frontal face. Among landmark-based loss function, we used L_2 loss to calculate the squared difference between the synthesized facial landmark and ground truth facial landmark coordinates.

L_2 loss expressed as:

$$L_2 = \frac{1}{N}\sum_i (F_i - F'_i)^2 \qquad eq\ 4.2$$

Where N is the total number of images. F_i and F'_i are the ground truth frontal image and synthesized frontal images respectively.

4.5.2 Multi-scale Pixel-wise Loss

In this study, the multi-scale pixel-wise loss introduced by (P. Li et al., n.d.-b) is employed to train both the generator and discriminator. This ensures the consistency of content between the synthesized (F') and the ground truth frontal facial images (F).

Multi-scale pixel wise loss expressed as:

$$L_{pixel} = \frac{1}{N}\sum_{n=1}^{N} \frac{1}{W_n H_n C} \sum_{w,h,c=1}^{W_n,\, H_n,\, C} \left| F'_{n,w,h,c} - F_{n,w,h,c} \right| \qquad eq\ 4.3$$

Where C is the number of channels, N is the number of scales. W_n , H_n represents the corresponding width and heights of scale n.

4.5.3 Total Variation Regularization Loss

We utilized total variation regularization loss (Johnson et al., 2016) to allow the model to produce smoother and visually appealing frontalized images, by reducing noise or artifacts while preserving important facial image features.

Total variation regularization loss expressed as:

$$L_{tv} = \sum_{c=1}^{C} \sum_{w,h=1}^{W,H} \mid F'_{w+1,h,c} - F'_{w,h,c} \mid + \mid F'_{w,h+1,c} - F'_{w,h,c} \mid \qquad eq\ 4.4$$

Where C is the number of channels, W is width of the generated frontal facial image, and H is height of the generated frontal facial image F'.

4.6 Training Pseudo Code

In this section, the training algorithm which used in this study is described.

Algorithm 1: Overall training algorithm of FFRAD-GAN model

1. **Initialize** *all generator G and discriminator D weight*

   Channel attention *at encoder and* ***self-attention*** *at decoder part of U-net*

   Relative average discriminator *with in discriminator part*
2. **Input***: batch of side-view (I_s) and frontal facial images (I_f)*
3. ***For*** *number of epochs* ***do***
4. ***For*** *number of batches in epoch* ***do***
5. ***Update*** *discriminator D using*

   $$Ep_{data}\left[\log\big(D(x)\big)\right] + Ep_z\left[\log\left(1 - D(G(z))\right)\right]$$
6. ***Update*** *generator G using*

   $$Ep_{data}\left[\log\big(D(x)\big)\right] + Ep_z\left[\log\Big(1 - D\big(G(z)\big)\Big)\right]$$

   $$\frac{1}{N}\sum_{i}(F_i - F'_i)^2$$
7. ***Save*** *the generator and discriminator weights*
8. ***End for***
9. **Output***: Frontalized frontal facial image*

   Bold, indicates the updated components

In above pseudocode x indicates a real data sample from the true data distribution p_{data}, $D(x)$ indicates the output of the discriminator when given a real data sample x, z indicates a noise vector sampled from the noise distribution p_z and $G(z)$ indicates the output of the generator when given the noise vector z. F_i also indicates ground truth image pixel and F'_i indicates synthesized image pixel.

CHAPTER FIVE

5. IMPLEMENTATION DETAILS

5.1 Chapter Overview

This chapter discusses about show the proposed model is implemented, along with the operational setting, techniques for data augmentation, the implementation of FFRAD-GAN, and the experimental procedures carried out.

5.2 Working Environment

This section explores the hardware tool setup used to execute the model within the environment.

Laptop:

- Processor: Intel(R) Core(TM) i5-4210M CPU @ 2.60GHz 2.60 GHz
- RAM: 4.00 GB
- Operating system: Window 10
- System type: 64-bit operating system, x64-based processor

Desktop:

- Processor: Intel(R) Core(TM) i7-8700M CPU @ 3.20GHz 3.19 GHz
- Installed RAM: 8.00 GB
- Operating system: Window 10
- System type: 64-bit operating system, x64-based processor
- Graphics: Graphics: Intel ® HD Graphics 520/ NVIDIA Corporation
- GPU: GeForce RTX 2070 Super 8 GB RAM

5.3 Environmental Setup

In this research, we explore various software programs and IDEs that support popular machine learning frameworks like TensorBoard and PyTorch, enabling us to effectively develop and analyze our models.

Anaconda: It is one of the working environments used in face frontalization using combination of attention and relative average based discriminator in GAN. It is a popular open-source installation of programming languages for scientific computation, data analysis, and machine learning. It comes with a large number of pre-installed packages and libraries, including prominent ones like NumPy, pandas, SciPy, Matplotlib, and scikit-learn, among others. The proposed solution utilizes Anaconda version 23.7.4 to install all essential libraries and modules needed for implementation.

Jupyter Notebook: It is well-suited for creating documented workflows and sharing them with others. The combination of code, explanations, and visualizations makes it easy for others to understand and reproduce the analysis.

Colab: is a cloud-based platform designed for writing and executing Python code. It eliminates the need for local setup, making it ideal for getting started quickly and provides free access to GPUs for a limited time. These powerful graphics processing units can significantly accelerate the execution of computationally intensive machine learning tasks, especially those involving deep learning models.

5.4 Implementation of Proposed Model

The proposed FFRAD-GAN model comprises of one U-net based generator and relative average discriminator. The U-net based generator, as described in section 4.3, composed of two network blocks: an encoder block and a decoder block which include hybrid attention mechanisms. Both the encoders block and the decoder block consist of six-layer neural networks with skip connections. Channel attention mechanism is utilized in encoder block, while self-attention mechanism is utilized in decoder block. ReLU activation function is used in both blocks of generator.

Table 5. 1: Content of encoder block in generator architecture

Architecture of Encoder block in Generator	
Layers in encoder	**Content of each block**
1	Convolution – (filters = 64, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True), Batch normalization, ReLU ()
2	Convolution – (filters = 128, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True), Batch normalization, ReLU ()
3	Convolution – (filters = 256, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True), Batch normalization, ReLU ()
4	Convolution – (filters = 512, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True), Batch normalization, ReLU ()
5	Convolution – (filters = 512, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True), Batch normalization, ReLU ()
6	Convolution – (filters = 512, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True), Batch normalization, ReLU ()

Table 5.1 describes about the contents used in encoder block of generator. There are six convolutional layers utilized in encoder block with kernel size of 4*4, stride of two and padding of one. We also used Batch normalization to stabilize the training of the generator to ensure that the generated samples have controlled distributions. ReLU activation function is also used in each layer to introduce non-linearity into the model and avoid vanishing gradients problem.

Table 5. 2: Content of bottleneck layer

	Content of Bottleneck layer
Bottleneck layer	Convolution – (filters = 512, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True), Batch normalization, ReLU () Deconvolution – (filters = 1024, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True), Batch normalization, ReLU ()

The bottleneck layer in a U-Net architecture plays a crucial role in bridging the encoder and decoder paths while capturing and transforming the deepest features extracted from the input data. As presented in table 5. 3, The bottleneck layer has two blocks: the convolutional and deconvolutional block. Both blocks have kernel size of 4*4, stride of two and padding of one. Batch normalization and ReLU activation function are utilized on both block of bottleneck layer.

Table 5. 4: Content of decoder block in generator architecture

Architecture of Decoder block in Generator	
Layers in decoder	**Content of each block**
1	Deconvolution – (filters = 1024, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True), Batch normalization, ReLU ()
2	Deconvolution – (filters = 1024, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True), Batch normalization, ReLU ()
3	Deconvolution – (filters = 1024, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True), Batch normalization, ReLU ()
4	Deconvolution – (filters = 512, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True), Batch normalization, ReLU
5	Deconvolution – (filters = 256, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True), Batch normalization, ReLU ()
6	Deconvolution – (filters = 128, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True), Tanh ()

Table 5.3 describes about the contents used in decoder block of generator. There are six deconvolutional layers utilized in decoder block with kernel size of 4*4, stride of two and padding of one. We also used Batch normalization in all layers except the last layer to stabilize the training of the generator to ensure that the generated samples have controlled distributions. ReLU activation function is also used in each layer except the last layer to introduce non-linearity into the model and avoid vanishing gradients problem. Due to our image is normalized in range of [-1,1], we used Tanh activation function in the last layer to output values in the range of [-1,1].

Table 5. 5: Content of relative average discriminator architecture

Layers in the RAD	**Content of each block**
1	Convolution – (filters = 16, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True), Batch normalization, LeakyReLU (0.2)
2	Convolution – (filters = 32, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True), Batch normalization, LeakyReLU (0.2)
3	Convolution – (filters = 64, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True), Batch normalization, LeakyReLU (0.2)
Average pooling	kernel size = (2,2), stride = (2,2)
4	Convolution – (filters = 128, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True), Batch normalization, LeakyReLU (0.2)
5	Convolution – (filters = 256, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True), Batch normalization, LeakyReLU (0.2)
6	Convolution – (filters = 512, kernel size = (4,4), stride = (2,2), padding = (1,1), bias = True)

As presented in table 5.4, the relative average discriminator contains two blocks with three convolutional layers in both of them with kernel size of 4*4, stride of two and padding one. Average pooling layer is utilized between two blocks to reduce the dimension of concatenated features by averaging the feature values. Batch normalization and LeakyReLU are used in each layer except the last layer and average pooling layer. In the last layer we utilized sigmoid activation function to produce a probability-like output that indicates the likelihood of the input image belonging to one of the two classes (real or fake).

5.4.1 Implementation of Self-attention

As stated in section 4.3.1 self-attention allows the generator to pay attention to not only fine details (local dependencies) but also how different parts of the face relate to each other across the entire image (global dependencies). This lets the generator create faces with much sharper details and finer structures, making them look more realistic. The self-attention layer is utilized at the third and fourth layers of the decoder.

Before applying self-attention, the input image is usually transformed into a sequence of vectors. Each vector (representing a pixel or a group of pixels) is treated as a query, a key, and a value. Query, key and value are obtained by convolving input image using kernel size of 1*1. For each query vector, dot products are computed with all key vectors to determine the importance (attention) of each key with respect to the query. This results in attention scores. These attention scores are then normalized using softmax to obtain attention weights. Finally, the weighted sum of value vectors, weighted by the attention scores, produces the output of the self-attention mechanism for each query. After the self-attention mechanism is applied to all vectors in the sequence, the resulting vectors are often reshaped back into the original spatial arrangement to retain spatial information.

5.4.2 Implementation of Channel Attention

Channel attention is utilized into the encoder block of the generator network to assist in focusing on relevant facial features.

In channel attention mechanism there are two blocks: squeezing and excitation blocks. In squeezing block, global average pooling and global Max pooling are applied to the input feature maps to generate channel descriptors. In excitation block there are two full connected layers: dimensionality reduction and dimensionality restoration. First the channel descriptors from squeezing block are passed through a fully connected layer with a reduction ratio r, reducing the number of channels from C to C/r. Next, the output of the first fully connected layer (i.e., dimensionality reduction) is passed through another fully connected layer that restores the number of channels back to C. Finally, a sigmoid activation function is applied to obtain the channel attention weights. The channel attention weights a are then used to reweight the original feature map X. Each channel of the feature map is scaled by the corresponding attention weight.

5.4.3 Implementation of Relative Average Discriminator

To enhance the discriminator's performance, we employed a relative average based discriminator for face frontalization in the discriminator part of the GAN. The RAD take two input images (i.e., real and fake image). Both of image are passed separately on subsequent convolutional layers to extract their features. After the features are extracted from both real and generated image, their features are combined together and passed to the average pooling layer to reduce dimension of concatenated features by averaging the feature values. Then, those features are passed to the next subsequent convolutional layers with sigmoid activation in last layers.

5.5 Hyperparameter Configuration

Hyperparameters are configuration settings used to control the learning process of a machine learning model. These variables are modifiable and directly influence the effectiveness of model training. Optimizers are methods that regulate a neural network's parameters throughout training to minimize the error or loss function. The weight of network was updated using the Adam optimizer with learning-rate=0.0002, beta_1=0.5, and beta_2=0.999. Adam combines the ideas of two other popular optimization algorithms, AdaGrad and RMSProp (Kingma & Ba, 2014). It provides automatic adjustment of learning rates for each parameter, and it is relatively easy to use with little hyperparameter tuning. The learning rate controls the step size throughout the optimization process, determining how much the model's parameters are adjusted with respect to the loss gradient. We trained the model with batch size of 32 for 100 epochs.

Table 5. 6: Hyperparameter configuration

Hyperparameter	Values
Batch size	32
Number of epochs	100
Optimizer	Adam
Learning Rate	0.0002

Table 5.5 presented the hyperparameters used during training our FFRAD-GAN model. Due to training GAN is computationally intensive, using large batch size is very difficult. And using very small batch size also increase the training time of the model. So, we selected using a moderate batch size (i.e., 32) considering the computational cost and training time of the model. We also selected Adam optimizer among another optimizers like SGD for its ability to adapts the learning rate for each parameter individually. This is beneficial for GANs as different parameters can converge at different rates, and an adaptive learning rate helps in stabilizing the training and faster convergence.

5.6 Experimental Class

To figure out the effectiveness of incorporating hybrid attention mechanisms and the relative average discriminator in the improvement of frontalized facial image quality, multiple experiments need to be conducted and properly tested.

There are five separate experiments carried out as presented in table 5.6. The initial experiment is presented as implementing the baseline work. As described in section 3.5, DMA-GAN utilized attention mechanisms and residual blocks with a deep feature encoder within generator part of the network. A conventional discriminator was employed for the discriminator component of the generative adversarial network. The second experiment serves as the starting experiment of the proposed model. We utilized self-attention block in decoder part of U-net based generator in this experiment. In the third experiment, we incorporated a channel attention block in the encoder part of the generator. In the fourth experiment, we utilized self-attention in generator and a relative average discriminator instead of the standard discriminator at the discriminator part of the GAN.

The fifth experiment introduces the comprehensive proposed model, which incorporates channel attention in the encoder sections and self-attention mechanisms in the decoder section of the U-net based generator network. Additionally, it utilizes the relative averages discriminator in the discriminator network, integrating multi-scale pixel-wise loss with adversarial loss.

Table 5. 7: List of experiment class

Notation	Experiment	Dataset used
DMA-GAN	Baseline	CAS-PEAL-R1 dataset
SA + DMA-GAN (**FFRAD-GAN-1**)	With utilizing self-attention at decoder parts of U-net based generator	CAS-PEAL-R1 and Multi-PIE datasets
SA + CA + DMA-GAN (**FFRAD-GAN-2**)	With utilizing channel attention at the encoder part and self-attention at decoder part of U-net based generator	CAS-PEAL-R1 and Multi-PIE datasets
SA + RAD + DMA-GAN (**FFRAD-GAN-3)**	With using self-attention in the decoder part of generator and relative average discriminator in discriminator part	CAS-PEAL-R1 and Multi-PIE datasets
SA + CA + DMA-GAN +RAD (**FFRAD-GAN-4**)	By using channel attention at the encoder and self-attention in the decoder part of generator, and relative average discriminator instead of standard discriminator at the discriminator parts of GAN	CAS-PEAL-R1 and Multi-PIE datasets

In table 5. 8 different experiments which are conducted on this study are summarized. In the first experiment, we utilized self-attention in decoder part of generator to solve the problem of blurriness and lacking fine details of images occurred when frontalizing images under larger pose angle above 45 ˚. Self-attention in the decoder allows the model to capture long-range dependencies and relationships within the image, which allows for preserving fine details and reducing blurriness.

In the second experiment, we utilized channel attention in encoder and self-attention in decoder part of generator to solve the problem of recovering facial attributes like eyeglass, hair and skin texture. Channel attention mechanisms enhance the model's ability to focus on the most informative feature channels, improving the extraction of important details such as eye-glasses, hair, and skin texture.

Another experiment is conducted with utilizing relative average discriminator to solve problems of lacking extraction of spatial context features and blurriness of frontalized image. The relative average discriminator improves the adversarial training process by providing more stable and meaningful gradients to the generator. This helps the generator in generating higher-quality images with better detail and less blurriness.

CHAPTER SIX

6. RESULTS AND DISCUSSION

6.1 Overview of Chapter

This chapter presents the results of the baseline work and the various experiments conducted, which include the new method we suggested. It presents the findings in an easy-to-understand way with charts, graphs, and tables, allowing for comparisons to be made.

6.2 Experimental Results

We trained the proposed model's network on multiple iterations with different factors. Five experiments were conducted utilizing identical datasets, hardware setups, and software configurations to ensure uniform comparison in order to guarantee reliable results for this study. We compared the effectiveness of our approach on two separate datasets: CAS-PEAL-R1 dataset (Wen Gao et al., 2008) and Multi-PIE dataset (P. Li et al., n.d.-a). The proposed FFRAD-GAN provides the best quantitative values based on quantitative assessment metrics and frontalizing faces with its features like hair, skin and also eye glass, as demonstrated by a comparison of our results with the state-of-the-art method.

6.3 Results based on Evaluation Metrics

To assess the model's performance objectively, we employed three metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Rank1-recognition rate. Based on (Tu et al., 2022), we used Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure Metrics (SSIM) to evaluate our model's performance. Due to PSNR measures how well the generated frontal view preserves the general structure and face details of original frontal facial image. And, beyond only evaluating noise levels, SSIM evaluates how well the generated frontal view accurately captures the spatial details of the original face, such as the locations of the eyes, the contours of the nose, and the lips. So, we got those metrics are an essential metrics to evaluate the proposed model's performance. Table 6.1 and 6.2 show the PSNR and SSIM, and rank1-recognition rate result of experiments on Multi-PIE dataset respectively. Table 6.3 and 6.4 also show PSNR and SSIM, and rank1-recognition rate result of experiments on CAS-PEAL-R1 dataset respectively. The better the PSNR, SSIM, and Rank-1 recognition rate, the more similar the generated frontal faces are to the originals in terms of quality and consistency

6.3.1 Evaluation of Experiments on Multi-PIE dataset

Table 6. 1: PSNR and SSIM evaluation on Multi-PIE dataset

S.NO	Experiments	PSNR	SSIM
1	DMA-GAN + SA (**FFRAD-GAN-1**)	36.95	0.9786
2	DMA-GAN + SA + CA (**FFRAD-GAN-2**)	37.71	0.9966
3	DMA-GAN+ SA +RAD (**FFRAD-GAN-3**)	37.86	0.9973
4	DMA-GAN+ SA+CA +RAD (**FFRAD-GAN-4**)	**38.76**	**0.9969**

To identify the best result in experiment, we've highlighted the best results in bold.

As a result, presented on table 6.1, our proposed FFRAD-GAN (i.e., FFRAD-GAN-4) approach outperforms all other approaches with Multi-PIE dataset image by improving PSNR values from 36.95 to **38.76** and SSIM value 0.9786 to **0.996**.

Table 6. 2: Rank-1 recognition rate evaluation on Multi-PIE dataset

Experiments	±15°	±30°	±45°	±60°	±75°	±90°	Avg
FFRAD-GAN-1	98.10	97.71	97.14	93.76	89.21	79.11	92.51
FFRAD-GAN-2	98.51	98.26	98.01	95.13	90.42	80.14	93.41
FFRAD-GAN-3	98.70	98.21	97.92	95.78	89.86	80.31	93.50
FFRAD-GAN-4	**99.89**	**99.83**	**99.31**	**97.73**	**94.65**	**82.92**	**95.72**

To identify the best result in experiment, we've highlighted the best results in bold. The value in above table is expressed by percentage (%).

As a result, presented on table 6.2, our proposed FFRAD-GAN (i.e., FFRAD-GAN-4) model achieves a greater rank-recognition rate result compared with other experiments by achieving **95.72%** of average value of all pose angles.

6.3.2 Evaluation of Experiments on CAS-PEAL-R1 dataset

Table 6. 3: PSNR and SSIM evaluation on CAS-PEAL-R1 dataset

S.NO	Experiments	PSNR	SSIM
1	**FFRAD-GAN-1**	35.34	0.9531
2	**FFRAD-GAN-2**	36.12	0.962
3	**FFRAD-GAN-3**	35.79	0.9542
4	**FFRAD-GAN-4**	**37.47**	**0.9765**

To identify the best result in experiment, we've highlighted the best results in bold.

FFRAD-GAN (FFRAD-GAN-4) achieves the highest PSNR and SSIM result for frontalization of side-view facial images, as shown in table 6.3, indicating that it outperforms other approaches.

Table 6. 4: Rank-1 recognition rate evaluation on CAS-PEAL-R1 dataset

Pitch	0°				30°			
Yaw	**±15°**	**±30°**	**±45°**	**Avg**	**±15°**	**±30°**	**±45°**	**Avg**
Experiments								
DMA-GAN	100.00	99.89	97.16	99.02	98.98	96.59	93.18	96.25
FFRAD-GAN-1	99.21	98.78	97.13	98.37	97.51	95.63	93.01	95.38
FFRAD-GAN-2	99.36	99.01	97.21	98.53	99.11	97.14	93.48	96.58
FFRAD-GAN-3	99.33	98.81	97.25	98.50	98.87	96.84	93.22	96.31
FFRAD-GAN-4	99.51	99.13	**97.42**	98.69	**99.24**	**97.19**	**94.10**	**96.84**

To identify the best result in experiment, we've highlighted the best results in bold. The value in table is expressed by percentage (%).

As shown in table 6.4, FFRAD-GAN (i.e., FFRAD-GAN-4) outperforms all other methods, achieving Rank1-recognition rate value from DMA-GAN results 98.98 to **99.24** on 15°, 96.59 to **97.19** on 30° and 93.18 to **94.10** on 45°.

In order to evaluate and interprate the performance of our model in accordance with the baseline work model, we conducted statistical analysis with alpha value of 0.05 and standard deviation of 0.5.

Based on the statistical analysis of the Rank1-recognition rates for face frontalization at different angles, we could draw conclusions regarding the performance of our model compared to the baseline model. At 15°, the difference in Rank1-recognition rate between our model and the baseline was 0.26, with a z-score of 0.65 and a p-value of 0.51569. Given that the p-value exceeds the alpha threshold of 0.05, indicating that there is no statistically significant difference between our model and the baseline at this angle.

Similarly, for the 30° pose, the difference in Rank1-recognition rate was 0.60, with a z-score of 1.50 and a p-value of 0.13361. Again, the p-value is greater than 0.05. This suggests that, at 30°, there is no significant difference in performance between our model and the baseline model.

However, at the 45° pose, the difference in Rank1-recognition rate is 0.92, with a z-score of 2.30 and a p-value of 0.02145. In this case, the p-value is less than the alpha threshold of 0.05. Therefore, we can conclude that at a 45° pose, there is a statistically significant difference between our model and the baseline model, indicating that our model performs significantly better at this angle.

Overall, while our model does not show a significant improvement over the baseline at 15° and 30°, it demonstrates a clear advantage at 45°. This suggests that our model's enhancements are particularly effective at larger angles of face rotation.

6.4 Frontalized Image Results of Experiments

In this section, the frontalized facial image are presented for all experiments including the baseline work. The first row indicates an input non-frontal facial image with their pose angle, the next rows are indicating the frontal results for the experiments and the last row indicate ground truth frontal-facial image.

6.4.1 Frontalized Results of Experiments for CAS-PEAL-R1 dataset

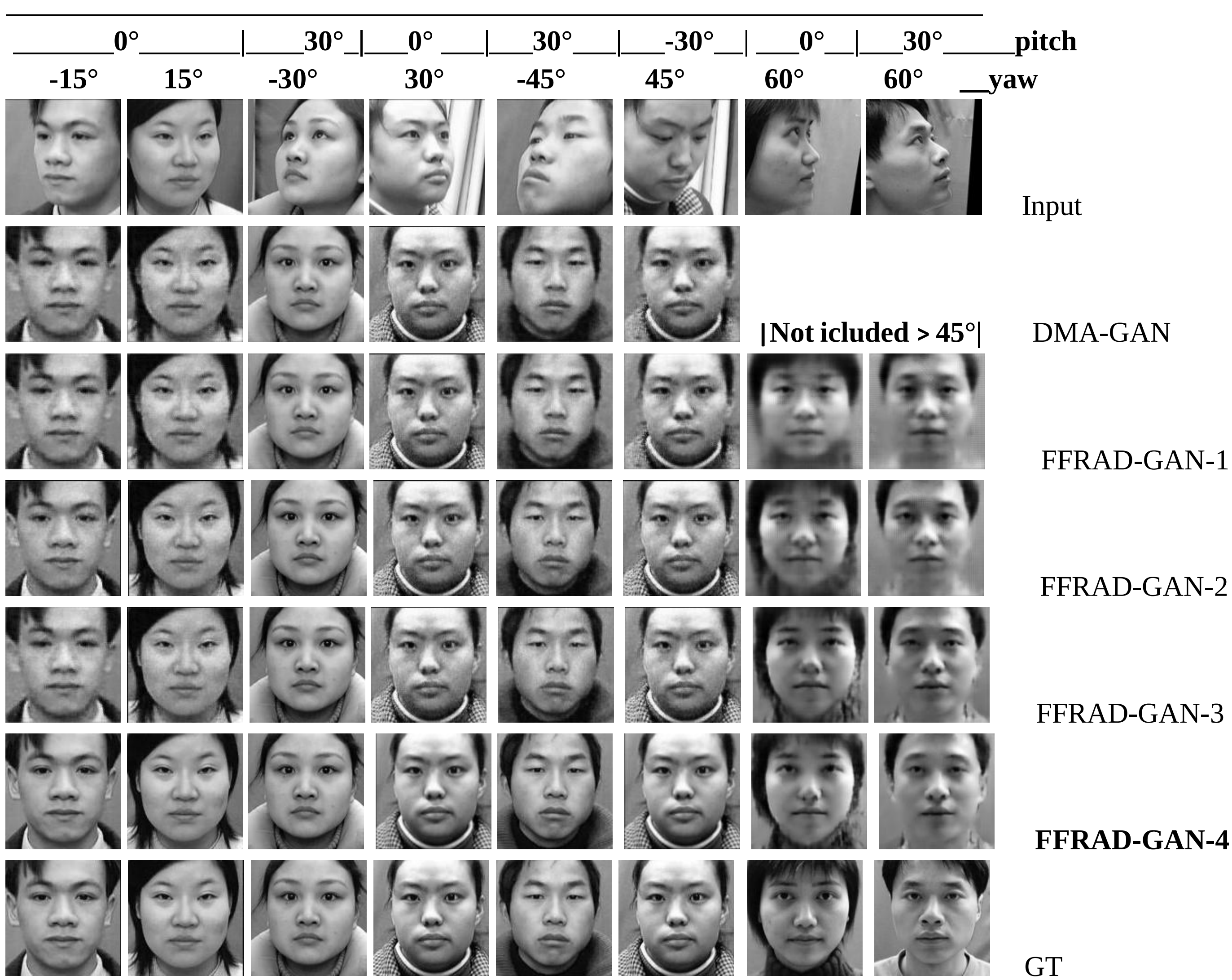

Figure 6. 1: CAS-PEAL-R1 frontalized facial images in different experiment classes.

In the figure above the first row indicates the pitch angle (i.e. upward and downward pose angle of 0° and ±30°) and the second row indicates the yaw angle (i.e. left and right pose angle: {±15°, ±30°, ±45°}). In the fourth row, Not included > 45° indicates that a pose angle greater than 45° is not included in baseline work (i.e., DMA-GAN).

6.4.2 Frontalized Results of Experiments for Multi-PIE dataset

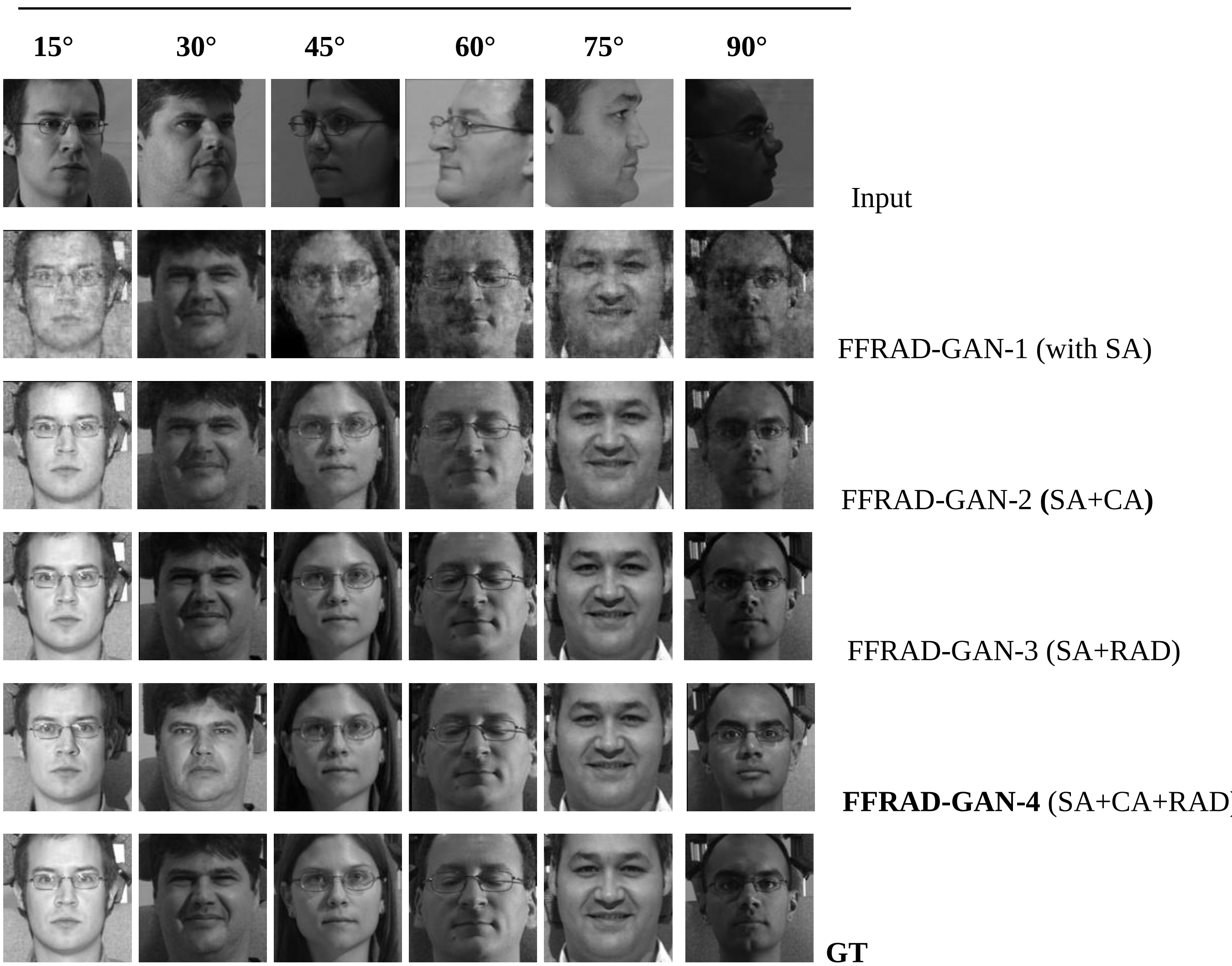

Figure 6. 2: Multi-PIE frontalized facial images in different experiment classes

In the depicted figure 6.2, the initial row represents the pose angle of the face, followed by the subsequent row displaying the input non-frontal facial image. The subsequent rows illustrate the frontalized face outcomes for the respective experimental classes. The last row shows the ground truth frontal image

6.4.3 DMA-GAN + SA (FFRAD-GAN-1)

To address the challenge of preserving structural details, a self-attention block was integrated into the decoder segment of the generator.

Figure 6. 3: Results of DMA-GAN + SA

In the figure 6.3, the first row indicates an input non-frontal facial image 0° to 90° and the next row indicates the corresponding frontal facial image frontalized by DMA-GAN + SA experimental class.

As shown in figure 6.3, the images are frontalized to its canonical view, but there is some blur on synthesized frontal image especially when the pose angle is increased.

6.4.4 DMA-GAN + SA + CA (FFRAD-GAN-2)

The second experiment was conducted by utilizing channel attention block in the encoder part of the generator network in addition to self-attention in decoder block to solve blurriness of synthesized frontal facial images.

Figure 6. 4: Results of DMA-GAN + SA + CA

In the figure 6.4, the first row indicates an input non-frontal facial image 0° to 90° and the next row indicates the corresponding frontal facial image frontalized by DMA-GAN + SA +CA expeimental class.

Based on figure 6.4, the model significantly reduces the blurriness as compared to the previous model results.

6.4.5 DMA-GAN + SA + RAD (FFRAD-GAN-3)

The third experiment was conducted by utilizing self-attention in the decoder block of the generator and relative average based discriminator in discriminator part of the network.

Figure 6. 5: Results of DMA-GAN + SA +RAD

In the figure 6.5, the first row indicates an input non-frontal facial image 0° to 90° and the next row indicates the corresponding frontal facial image frontalized by our proposed FFRAD-GAN experiment.

6.4.6 DMA-GAN + SA + CA +RAD (FFRAD-GAN-4)

In this experiment, we utilized channel attention block in the encoder part, self-attention block in the decoder part of generator network. Instead of standard discriminator, we used relative average discriminator in order to enhance performance of the discriminator.

Figure 6. 6 : Results of DMA-GAN + SA + CA + RAD (FFRAD-GAN)

In the figure 6.6, the first row indicates an input non-frontal facial image 0° to 90° and the next row indicates the corresponding frontal facial image frontalized by our proposed FFRAD-GAN experiment.

Our model can frontalizes non-frontal facial images to its frontal view and reduces blurriness on all pose angles from 15° up to 90°.

6.5 Sample Training Log for FFRAD-GAN

We used charts and graphs (training log figures) to monitor the model's performance as it learned. These figures tracked different aspects of the training process, like how well the model minimized errors (loss), and offer vital information about the model's effectiveness and training status for frontalizing side-view facial images.

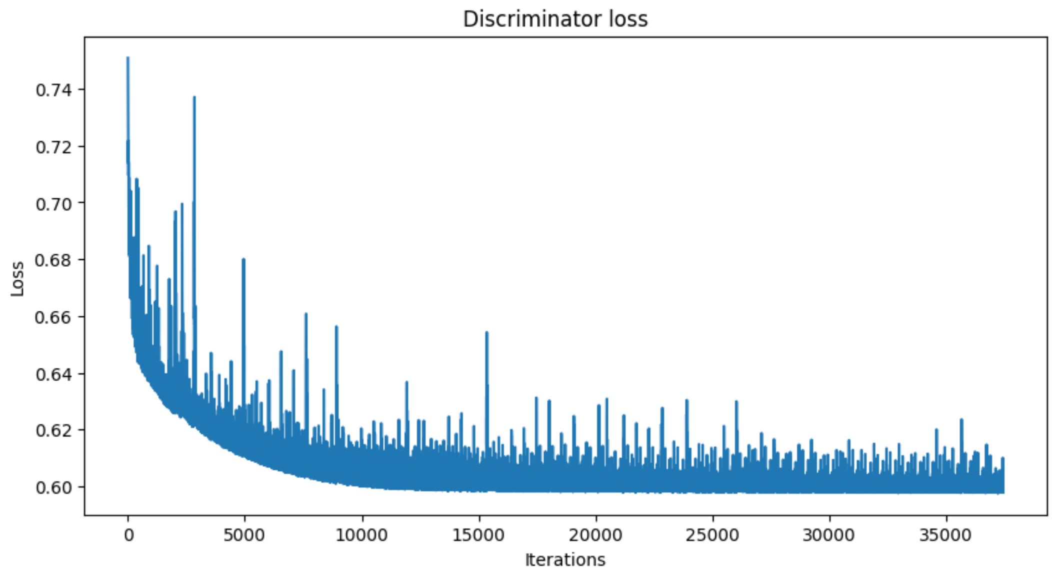

Figure 6. 7: Discriminator loss during training on CAS-PEAL-R1 dataset

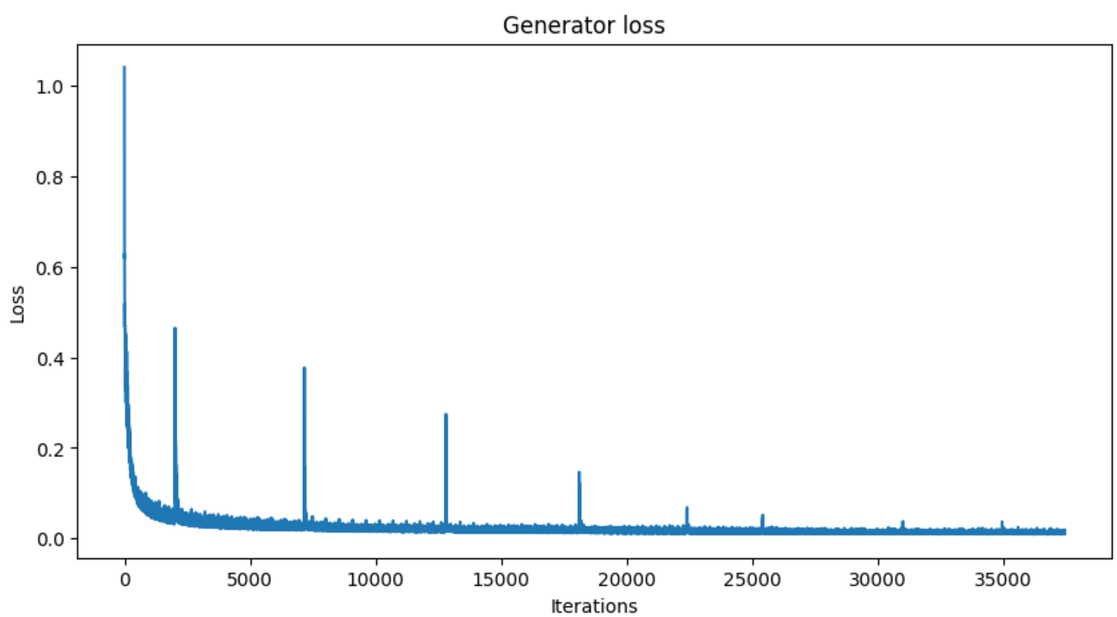

Figure 6. 8: Generator loss during training on CAS-PEAL-R1 dataset

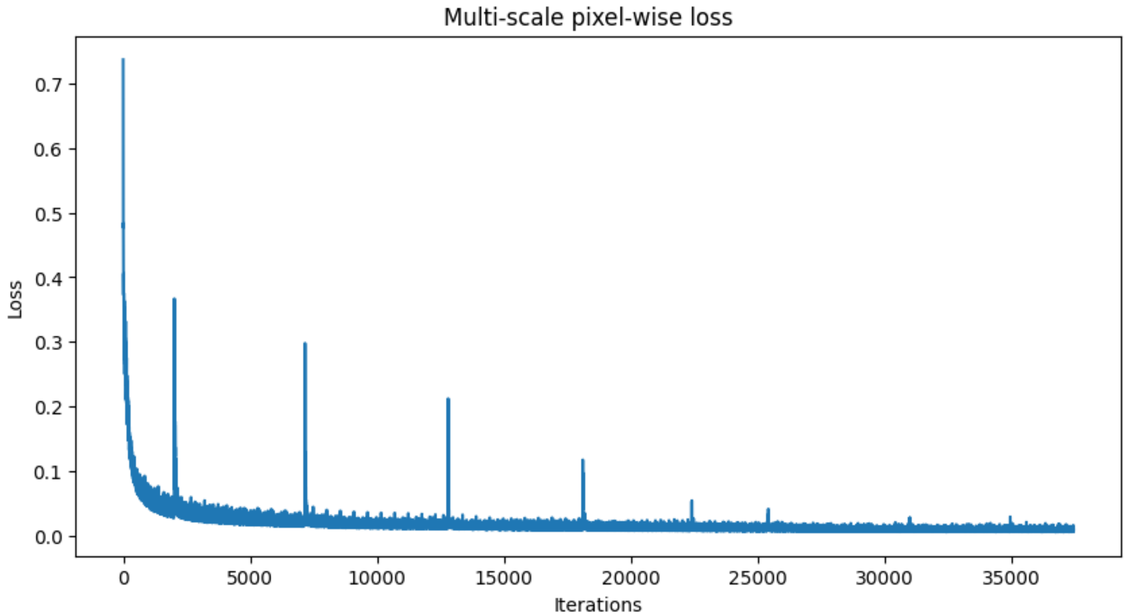

Figure 6. 9: Multi scale pixel-wise loss during training on CAS-PEAL-R1 dataset

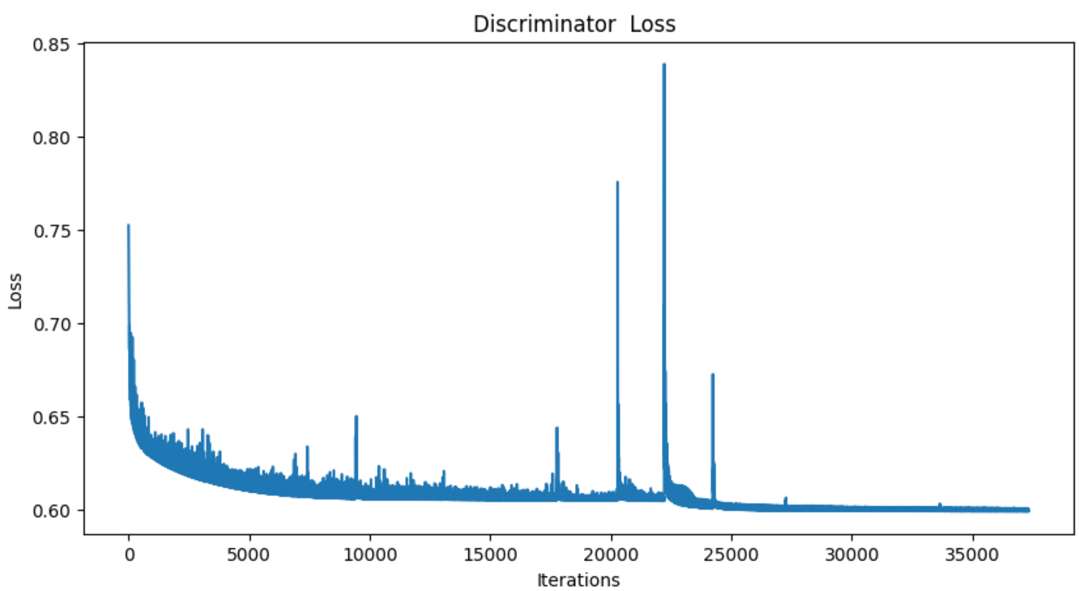

Figure 6. 10: Discriminator loss during training on Multi-PIE dataset

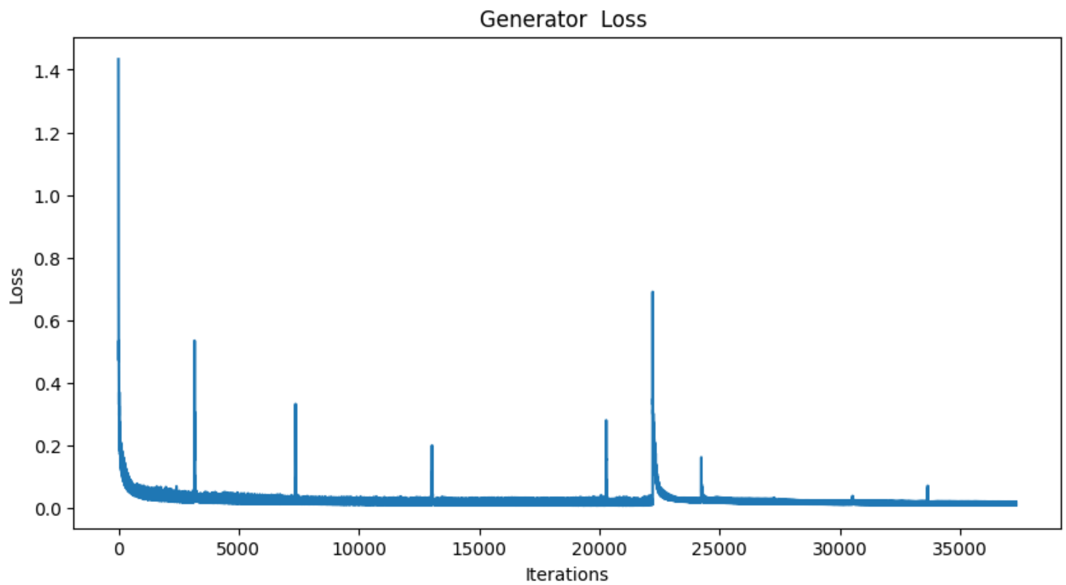

Figure 6. 11: Generator loss during training on Multi-PIE dataset

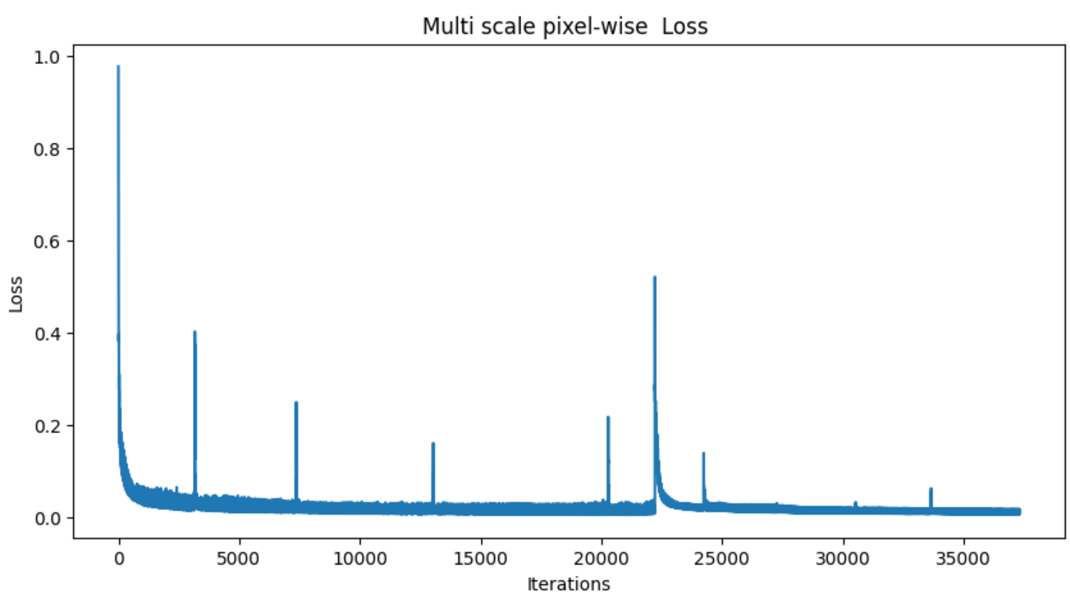

Figure 6. 12: Multi scale pixel-wise loss during training on Multi-PIE dataset

The study's findings suggest that the FFRAD-GAN method is successful in converting side-view facial images into a canonical frontal view. U-net, relative average discriminator, self-attention, and channel attention were combined by FFRAD-GAN. Based on the various quantitative metrics, the performance of FFRAD-GAN is better than another method like GSP-GAN (Luan et al., 2020), DA-GAN (Yin et al., 2020) and DMA-GAN (Cao et al., 2023). The self-attention (H. Zhang et al., 2018) and channel attention (Hu et al., 2018) improved the capability of generator and, relative average discriminator (Jolicoeur-Martineau, 2019) improved the performance of discriminator of the model. Using FFRAD-GAN, the frontalized facial image details and features are synthesized effectively.

6.6 Research Question and Answer Discussion

Answer for Q1: This work utilized a self-attention and channel attention in U-net based generator, and a relative average based discriminator in discriminator part of the network. Based on experimental results, there was improvement on the quality of frontalized facial images even when the pose is become larger. As demonstrated by PSNR, SSIM, and Rank-1 recognition rate based on FFRAD-GAN on table 6.1, table 6.2 for Multi-PIE and, table 6.3, table 6.4 for CAS-PEAL-R1 dataset, channel attention, self-attention mechanisms and a relative average based discriminator is effective in increasing the quality of frontalized facial images and frontalizing face with its fine features. So, this study concludes that incorporating channel attention, self-attention, and a relative average based discriminator (RAD) with in GAN significantly improves the quality of frontalized facial images.

Answer for Q2: The optimal learning constraints to enhance the quality of frontalized facial images were observed. These include employing channel attention and self-attention to enhance structural details, utilizing multi-scale loss to compare pixels of the generated frontal image with the corresponding pixels of the real frontal image and using total variation regularization loss to reduce noise or artifacts on synthesized frontal facial image.

CHAPTER SEVEN

7. CONCLUSION AND FUTURE WORKS

7.1 Conclusion

Recent advancements in Generative Adversarial Networks (GANs) have significantly boosted computer vision tasks, including face frontalization. This paper proposes an approach for this task, leveraging a generative adversarial network that incorporates a combination of attention mechanism and a relative average based discriminator. Generating high-quality faces which include the face attributes like eyeglass, skin, and hair is a challenging task especially in extreme poses above 45˚. In previous methods like DMA-GAN, there are limitations when frontalizing side-view facial images and the images tends to be blur when the pose angle is being larger. Generally, in this study we presented a method for a face fronalization task using a generative adversarial network. The generator network utilized U-net architecture designed with incorporating combination of attention, trained adversarial with relative average discriminator to increase the quality and effectiveness of face frontalization task.

Enhancing the model's performance for face frontalization tasks is the main objective of this study. To make model effective, we proposed FFRAD-GAN, in which channel attention is used in encoder part, self-attention in decoder block of U-net generator, and relative average discriminator as a discriminator. This approach can effectively frontalize non-frontal (side-view) facial images into its canonical frontal view even the pose becomes larger. Using self-attention and channel attention increased the model's performance significantly. This study also used multi-scale pixel wise loss and total variation regularization loss in order to compare the pixels of the generated (frontalized) facial image with the corresponding pixels of the real frontal facial image, and to produce smoother and visually appealing frontalized images by reducing noise or artifacts.

Different experiments were conducted to make comparisons among each other. The first experiment is DMA-GAN + SA which utilized self-attention in decoder part of U-net based generator which made improvement on the generator compared to DMA-GAN.

The second experiment is DMA-GAN + SA +CA in which channel attention is used in encoder part of generator. It shows improvement on the performance of the generator as compared to DMA-GAN.

The third experiment is FFRAD-GAN which used relative average discriminator in discriminator part of the network. The performance of discriminator is improved by using relative average discriminator compared to a standard discriminator which utilized in DMA-GAN. FFRAD-GAN is an effective method for frontalizing face with high quality and finer features among all other experiments based on rank1 recognition rate, PSNR, and SSIM values.

Our experiments evaluated the FFRAD-GAN method using different quantitative measures. The results demonstrate that FFRAD-GAN outperforms existing facial frontalization techniques with quality of generated image and frontalized facial image features. In general, the results show that the FFRAD-GAN approaches are effective for face frontalization tasks including frontalization of face with larger pose angles. These findings have significant for face frontalization applications. As a result, we are confident that implementing FFRAD-GAN holds considerable potential within face recognition, security, and surveillance systems.

7.2 Future Works

This study aimed to enhance transformation of side-view facial images into its canonical frontal facial images, which are crucial for enhancing the performance of face recognition technologies. As a future work, we suggest integrating the FFRAD-GAN model to serve as a real-time face frontalizer within face recognition systems. Our next step involves evaluating the efficacy of combining attention and relative average based discriminator in GAN models for facial frontalization in the context of face recognition technologies.

Another direction for future research in face frontalization tasks is to explore the utilization of unsupervised or self-supervised learning approaches. This approach utilizes vast amounts of unlabeled data, potentially leading to models that generalize better. They improve adaptability of frontalization models across diverse facial variations and conditions, thereby enhancing their effectiveness in real-world face recognition scenarios.

SPECIAL ACKNOWLEDGMENT

This research work was funded by grant number ASTU/SM-R/1063/24 from Adama Science and Technology University.

REFERENCES

Al-Azzawi, A., Hind, J., & Cheng, J. (2018). Localized Deep-CNN Structure for Face Recognition. *2018 11th International Conference on Developments in ESystems Engineering (DeSE)*, 52–57. https://doi.org/10.1109/DeSE.2018.00049

Asthana, A., Marks, T. K., Jones, M. J., Tieu, K. H., & Rohith, M. (2011). Fully automatic pose-invariant face recognition via 3D pose normalization. *2011 International Conference on Computer Vision*, 937–944. https://doi.org/10.1109/

ICCV.2011.6126336

Cao, J., Chen, Z., Zhang, Y., Sun, L., & Chen, J. (2023). Face frontalization with deep GAN via multi-attention mechanism. *Signal, Image and Video Processing*, *17*(5), 1965–1973. https://doi.org/10.1007/s11760-022-02409-7

Cen, S., Luo, H., Huang, J., Shi, W., & Chen, X. (2022). Pre-Trained Feature Fusion and Multidomain Identification Generative Adversarial Network for Face Frontalization. *IEEE Access*, *10*, 77872–77882. https://doi.org/10.1109/ACCESS.2022.3193386

Chai, C., Liao, J., Zou, N., & Sun, L. (2018). A one-to-many conditional generative adversarial network framework for multiple image-to-image translations. *Multimedia Tools and Applications*, *77*(17), 22339–22366. https://doi.org/10.1007/s11042-018-5968-7

Chen, Y.-C., Shen, X., Lin, Z., Lu, X., Pao, I.-M., & Jia, J. (2019). Semantic Component Decomposition for Face Attribute Manipulation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9851–9859.

https://doi.org/10.1109/ CVPR.2019.01009

Ding, C., & Tao, D. (2015). Robust Face Recognition via Multimodal Deep Face Representation. *IEEE Transactions on Multimedia*, *17*(11), 2049–2058. https://doi.org/10.1109/TMM.2015.2477042

Ding, C., & Tao, D. (2017). Pose-invariant face recognition with homography-based normalization. *Pattern Recognition*, *66*, 144–152. https://doi.org/10.1016/j.patcog.

2016.11.024

Gauthier, J. (2015). *Conditional generative adversarial nets for convolutional face generation*. https://api.semanticscholar.org/CorpusID:3559987

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020a). Generative Adversarial Networks. *COMMUNICATIONS OF THE ACM*, *63*(11). https://doi.org/10.1145/3422622

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020b). Generative adversarial networks.

Communications of the ACM, *63*(11), 139–144. https://doi.org/10.1145/3422622

Hassner, T., Harel, S., Paz, E., & Enbar, R. (2015). Effective face frontalization in unconstrained images. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4295–4304. https://doi.org/10.1109/CVPR.2015.7299058

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition*.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation Networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132–7141.

https://doi.org/10.1109/CVPR.2018.00745

Huang, R., Zhang, S., Li, T., & He, R. (2017). Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2458–2467. https://doi.org/10.1109/ICCV.2017.267

Isa, I. S., Sulaiman, S. N., Mustapha, M., & Darus, S. (2015). Evaluating Denoising Performances of Fundamental Filters for T2-Weighted MRI Images. *Procedia Computer Science*, *60*, 760–768. https://doi.org/10.1016/j.procs.2015.08.231

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-Image Translation with Conditional Adversarial Networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5967–5976. https://doi.org/10.1109/CVPR.2017.632

Jeni, L. A., & Cohn, J. F. (2016). Person-Independent 3D Gaze Estimation Using Face Frontalization. *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 792–800. https://doi.org/10.1109/CVPRW.2016.104

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). *Perceptual Losses for Real-Time Style Transfer and Super-Resolution*.

Jolicoeur-Martineau, A. (2019). *THE RELATIVISTIC DISCRIMINATOR: A KEY ELEMENT MISSING FROM STANDARD GAN*. https://github.com/AlexiaJM/

RelativisticGAN.

Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Dusik Park, & Junmo Kim. (2015). Rotating your face using multi-task deep neural network. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 676–684. https://doi.org/10.1109/CVPR.2015.7298667

Kammoun, A., Slama, R., Tabia, H., Ouni, T., & Abid, M. (2023). Generative Adversarial Networks for Face Generation: A Survey. *ACM Computing Surveys*, *55*(5), 1–37. https://doi.org/10.1145/3527850

Kang, S., Lee, J., Bong, K., Kim, C., Kim, Y., & Yoo, H.-J. (2018). Low-Power Scalable 3-D Face Frontalization Processor for CNN-Based Face Recognition in Mobile Devices. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, *8*(4), 873–883. https://doi.org/10.1109/JETCAS.2018.2845663

Kingma, D. P., & Ba, J. (2014). *Adam: A Method for Stochastic Optimization*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, *1*(4), 541–551. https://doi.org/10.1162/neco.1989.1.4.541

Li, P., Wu, X., Hu, Y., He, R., & Sun, Z. (n.d.-a). *M 2 FPA: A Multi-Yaw Multi-Pitch High-Quality Dataset and Benchmark for Facial Pose Analysis*.

Li, P., Wu, X., Hu, Y., He, R., & Sun, Z. (n.d.-b). *M 2 FPA: A Multi-Yaw Multi-Pitch High-Quality Dataset and Benchmark for Facial Pose Analysis*.

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(12), 6999–7019. https://doi.org/10.1109/

TNNLS.2021.3084827

Liakos, K., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine Learning in Agriculture: A Review. *Sensors*, *18*(8), 2674. https://doi.org/10.3390/s18082674

Luan, X., Geng, H., Liu, L., Li, W., Zhao, Y., & Ren, M. (2020). Geometry Structure Preserving Based GAN for Multi-Pose Face Frontalization and Recognition. *IEEE Access*, *8*, 104676–104687. https://doi.org/10.1109/ACCESS.2020.2996637

Luo, H., Cen, S., Ding, Q., & Chen, X. (2022). Frontal face reconstruction based on detail identification, variable scale self-attention and flexible skip connection. In *Neural Computing and Applications* (Vol. 34, Issue 13, pp. 10561–10573). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s00521-022-07124-5

Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., & Smolley, S. P. (2017). Least Squares Generative Adversarial Networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2813–2821. https://doi.org/10.1109/ICCV.2017.304

Mirza, M., & Osindero, S. (n.d.). *Conditional Generative Adversarial Nets*.

Pang, Y., Lin, J., Qin, T., & Chen, Z. (2022). Image-to-Image Translation: Methods and Applications. *IEEE Transactions on Multimedia*, *24*, 3859–3881. https://doi.org/10.1109/TMM.2021.3109419

Parmar, D. N., & Mehta, B. B. (2014). *Face Recognition Methods & Applications*.

Perarnau, G., Van De Weijer, J., Raducanu, B., & Álvarez, J. M. (n.d.). *Invertible Conditional GANs for image editing*. Retrieved December 7, 2023, from https://github.com/Guim3/IcGAN

Ramzan, F., Khan, M. U. G., Rehmat, A., Iqbal, S., Saba, T., Rehman, A., & Mehmood, Z. (2020). A Deep Learning Approach for Automated Diagnosis and Multi-Class Classification of Alzheimer's Disease Stages Using Resting-State fMRI and Residual Neural Networks. *Journal of Medical Systems*, *44*(2), 37. https://doi.org/10.1007/s10916-019-1475-2

Rong, C., Zhang, X., & Lin, Y. (2020). Feature-Improving Generative Adversarial Network for Face Frontalization. *IEEE Access*, *8*, 68842–68851. https://doi.org/10.1109/ACCESS.2020.2986079

Ronneberger, O., Fischer, P., & Brox, T. (2015a). *U-Net: Convolutional Networks for Biomedical Image Segmentation*.

Ronneberger, O., Fischer, P., & Brox, T. (2015b). *U-Net: Convolutional Networks for Biomedical Image Segmentation*.

Rosenfeld, A. (1988). Computer vision: basic principles. *Proceedings of the IEEE*, *76*(8), 863–868. https://doi.org/10.1109/5.5961

Sagonas, C., Panagakis, Y., Zafeiriou, S., & Pantic, M. (2015). Robust Statistical Face Frontalization. *2015 IEEE International Conference on Computer Vision (ICCV)*, 3871–3879. https://doi.org/10.1109/ICCV.2015.441

Sagonas, C., Panagakis, Y., Zafeiriou, S., & Pantic, M. (2017). Robust Statistical Frontalization of Human and Animal Faces. *International Journal of Computer Vision*, *122*(2), 270–291. https://doi.org/10.1007/s11263-016-0920-7

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Shen, W., & Liu, R. (2017). Learning Residual Images for Face Attribute Manipulation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1225–1233. https://doi.org/10.1109/CVPR.2017.135

Shen, Y., Luo, P., Yan, J., Wang, X., & Tang, X. (2018). FaceID-GAN: Learning a Symmetry Three-Player GAN for Identity-Preserving Face Synthesis. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 821–830. https://doi.org/10.1109/CVPR.2018.00092

Shu, Z., Yumer, E., Hadap, S., Sunkavalli, K., Shechtman, E., & Samaras, D. (2017). Neural Face Editing with Intrinsic Image Disentangling. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5444–5453. https://doi.org/

10.1109/CVPR.2017.578

Shukla, S., Van Gool, L., & Timofte, R. (2019). Extremely Weak Supervised Image-to-Image Translation for Semantic Segmentation. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 3368–3377. https://doi.org/

10.1109/ICCVW.2019.00419

Thies, J., Zollhöfer, M., Nießner, M., Valgaerts, L., Stamminger, M., & Theobalt, C. (2015). Real-time expression transfer for facial reenactment. *ACM Transactions on Graphics*, *34*(6), 1–14. https://doi.org/10.1145/2816795.2818056

Tu, X., Zhao, J., Liu, Q., Ai, W., Guo, G., Li, Z., Liu, W., & Feng, J. (2022). Joint Face Image Restoration and Frontalization for Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(3), 1285–1298. https://doi.org/10.1109/

TCSVT.2021.3078517

Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience*, *2018*, 1–13. https://doi.org/10.1155/2018/7068349

Wang, L., Mostafavi, I. S. M., Ho, Y.-S., & Yoon, K.-J. (2019). Event-Based High Dynamic Range Image and Very High Frame Rate Video Generation Using Conditional Generative Adversarial Networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10073–10082. https://doi.org/10.1109/CVPR.2019.01032

Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, & Debin Zhao. (2008). The CAS-PEAL Large-Scale Chinese Face Database and Baseline Evaluations. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *38*(1), 149–161. https://doi.org/10.1109/TSMCA.2007.909557

Wenchao Zhang, Shiguang Shan, Gao, W., Xilin Chen, & Hongming Zhang. (2005). Local Gabor binary pattern histogram sequence (LGBPHS): a novel non-statistical model for face representation and recognition. *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, 786-791 Vol. 1. https://doi.org/10.1109/ICCV.2005.147

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). *CBAM: Convolutional Block Attention Module*.

Wu, Y., & Feng, J. (2018). Development and Application of Artificial Neural Network. *Wireless Personal Communications*, *102*(2), 1645–1656. https://doi.org/10.1007/s11277-017-5224-x

Yeh, R., Liu, Z., Goldman, D. B., & Agarwala, A. (2016). *Semantic Facial Expression Editing using Autoencoded Flow*.

Yin, X., Yu, X., Sohn, K., Liu, X., & Chandraker, M. (2017). Towards Large-Pose Face Frontalization in the Wild. *2017 IEEE International Conference on Computer Vision (ICCV)*, 4010–4019. https://doi.org/10.1109/ICCV.2017.430

Yin, Y., Jiang, S., Robinson, J. P., & Fu, Y. (2020). Dual-Attention GAN for Large-Pose Face Frontalization. *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, 249–256. https://doi.org/10.1109/ FG47880.2020.00004

Yuan, Z., Jiang, M., Wang, Y., Wei, B., Li, Y., Wang, P., Menpes-Smith, W., Niu, Z., & Yang, G. (2020). SARA-GAN: Self-Attention and Relative Average Discriminator Based Generative Adversarial Networks for Fast Compressed Sensing MRI Reconstruction. *Frontiers in Neuroinformatics*, *14*. https://doi.org/10.3389/ fninf.2020.611666

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2018). *Self-Attention Generative Adversarial Networks*.

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., & Fu, Y. (2018). *Image Super-Resolution Using Very Deep Residual Channel Attention Networks*.

Zhang, Z., Chen, X., Wang, B., Hu, G., Zuo, W., & Hancock, E. R. (2019). Face Frontalization Using an Appearance-Flow-Based Convolutional Neural Network. *IEEE Transactions on Image Processing*, *28*(5), 2187–2199. https://doi.org/10.1109/ TIP.2018.2883554

Zhang, Z., Liang, R., Chen, X., Xu, X., Hu, G., Zuo, W., & Hancock, E. R. (2021). Semi-Supervised Face Frontalization in the Wild. *IEEE Transactions on Information Forensics and Security*, *16*, 909–922. https://doi.org/10.1109/TIFS.2020.3025412

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks*.