CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD

Contrastive Unpaired Translation Network for Visible to Thermal (CUTV2T) Transformation of Facial Images

by

Usama Ahmad

A thesis submitted in partial fulfillment for the
degree of Master of Science

in the

Faculty of Engineering
Department of Electrical Engineering

2022

I would like to dedicate this thesis to my parents. To my father that supported me, taught me the value of education, and guided me to where I am today. To my mom that raised me, taught me the value of love, and was always there for me. Without them none of this would have been possible

CERTIFICATE OF APPROVAL

Contrastive Unpaired Translation Network for Visible to Thermal (CUTV2T) Transformation of Facial Images

by

Usama Ahmad

(MEE193027)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr.Abdul Ghafoor	MCS-NUST,Islamabad
(b)	Internal Examiner	Dr. M. Ashraf	CUST, Islamabad
(c)	Supervisor	Dr. Imtiaz Ahmad Taj	CUST, Islamabad

Dr. Imtiaz Ahmad Taj
Thesis Supervisor
November,2022

Dr. Noor Muhammad Khan
Head
Dept. of Electrical Engineering
November, 2022

Dr. Imtiaz Ahmad Taj
Dean
Faculty of Engineering
November, 2022

Author's Declaration

I, **Usama Ahmad** hereby state that my MS thesis titled "**Contrastive Unpaired Translation Network for Visible to Thermal (CUTV2T) Transformation of Facial Images**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

(Usama Ahmad)
Registration No: MEE193027

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**Contrastive Unpaired Translation Network for Visible to Thermal (CUTV2T) Transformation of Facial Images**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

(Usama Ahmad)
Registration No: MEE193027

Acknowledgement

Then which of the Blessings of your Lord will you deny. (Al-Quran).

First and foremost to the creator, the most gracious, the most beneficent, the Almighty ALLAH S.W.T, I owe it all to you, Thank you!

There have been many people who have walked alongside me, who have guided me through all these efforts. I would like to outstretch gratitude to each one of them.Topping the list is my supervisor Dr. Imtiaz Ahmad Taj to whom i owe my deepest gratitude for providing his valuable guidance to complete this research.Beside that i am also very grateful to my lab-mate Mr. Mohsin Ullah for his unconditional help as well as technical motivational support thorough out the research journey. Moreover, I would like to thank each member of the VisPRS research group for their kindness.

Furthermore, I owe a great deal to my teachers and parents who shaped me into the person I am today. Their continuous support and encouragement made this work possible. Nevertheless, I also want to acknowledge my mother's unconditional love and unending prayers for me.

(Usama Ahmad)

Abstract

Thermal to visible (T2V) image translation, or vice versa, is fundamentally an ambiguous problem because thermal image does not have any information about colors or brightness, or visible RGB image does not have any information about thermal signatures of different objects. However, with the advancement of deep learning this topic has attracted a lot of interest among researchers due to its important utility especially in surveillance applications. As compared to T2V translation, work on visible to thermal (V2T) translation models is very rare due to the fact that the prediction accuracy of such systems is mostly below par as sufficient training data to learn heat signatures of different parts of images is mostly not available. This study is motivated by this main limitation and the main goal is to develop a V2T translation model that can be trained despite lack of proper training data. Translation of facial image is targeted as facial images have a typical structure and symmetry not only according to visible features but also thermal signatures. A new model, namely, contrastive unpaired translation network for visible to thermal transformation (CUTV2T) is proposed for facial image translation that can be trained on unpaired visible and thermal facial images thus overcoming the main limitation of data scarcity.The proposed CUTV2T framework uses the CycleGAN [1] configuration except that the contrastive loss is replaced with the l_1 loss function. The CUTV2T selects two related patches (positives) compared to other patches (negatives) in the dataset and then maps both patches to the exact location in the learned features space. In the context of unpaired image-to-image translation, it is shown that the strategy used enables one-sided translation while enhancing the quality and lowering training time. Further, the performance of the CUTV2T model has been evaluated by determining evaluation parameters like FID score, SSIM, PSNR, and UQI values on two well-known databases Carl [2] and Tuft [3]. The results show that CUTV2T outperforms the CycleGAN [1] and pix2pix [4] models on both Carl and Tuft databases, thus demonstrating its efficacy for real-world applications.

Contents

Author's Declaration iv

Plagiarism Undertaking v

Acknowledgement vi

Abstract vii

List of Figures xi

List of Tables xiii

Abbreviations xiv

Symbols xv

1 Introduction 1
- 1.1 Motivations . . . 1
- 1.2 Purpose . . . 2
- 1.3 Background . . . 3
- 1.4 Structure of the Thesis . . . 4
- 1.5 Summary . . . 5

2 Literature Review 6
- 2.1 Introduction . . . 6
  - 2.1.1 Variational Autoencoder (VAEs) . . . 7
- 2.2 GANs Based Synthesis . . . 10
  - 2.2.1 GANs for Visible to Thermal Image Translation . . . 12
    - 2.2.1.1 Pix2Pix Based Approaches . . . 13
    - 2.2.1.2 CycleGAN . . . 14
    - 2.2.1.3 Variants of CycleGANs for Image Translation . . . 17
    - 2.2.1.4 Other GAN Architectures for Thermal Face Generation . . . 19
  - 2.2.2 Image Translation Network Architectures other than GANs . 21
    - 2.2.2.1 ThermelNet . . . 21
    - 2.2.2.2 Synthesis Based on Local Linearity . . . 22

2.2.2.3 Synthesis of Visible Faces Using Local Linear Regression . . . 22
2.2.2.4 Cascaded Refinement Network (CRN) . . . 22
2.3 Gap Analysis . . . 24
2.4 Problem Statement . . . 24
2.5 Research Contributions . . . 25
2.6 Summary . . . 25

3 Methodology **26**
3.1 Background . . . 26
3.2 Proposed Methodology . . . 27
3.2.1 Architectures . . . 28
3.2.1.1 Generator . . . 28
3.2.1.2 Discriminator . . . 29
3.2.2 Loss Functions . . . 30
3.2.3 Adversarial Loss . . . 30
3.2.4 Patchwise Contrastive Loss . . . 32
3.3 Summary . . . 34

4 Results and Discussion **35**
4.1 Datasets . . . 35
4.1.1 Carl Database . . . 35
4.1.2 Tuft Face Database . . . 37
4.2 Evaluation Metrics . . . 39
4.2.1 FID . . . 39
4.2.2 SSIM . . . 40
4.2.3 PSNR . . . 42
4.2.4 UQI . . . 42
4.3 Preprocessing with Retinaface . . . 43
4.4 Training Details . . . 44
4.4.1 Experiments with Different Parameter Settings. . . . 45
4.5 Discussion . . . 49
4.5.1 Taking Negatives from within the Same Image is more Powerful . . . 49
4.5.2 The Importance of Employing Multiple Encoder Layers . . . 50
4.5.3 The Regularizer $L_{PatchNCE}(G, H, Y)$ Stabilizes Training . . . 50
4.5.4 Updating Decoder without PatchNCE . . . 50
4.5.5 How does our Model Transform the Visible Face Image to the Thermal Face Image? . . . 51
4.6 Computational Resources . . . 51
4.7 Quantitative Comparisons . . . 51
4.8 Qualitative Comparisons . . . 52
4.9 Summary . . . 54

5 Conclusion and Challenges **55**
5.1 Conclusion . . . 55

5.2 Future Work 55
5.3 Challenges 56

Bibliography **57**

List of Figures

2.1 Encoder compresses data in the Latent Space (Z) 7
2.2 The Decoder Reconstruct the Data Given the Hidden Representation 8
2.3 Generative Adversarial Network's workflow 9
2.4 GAN loss function terms, from [35] 10
2.5 The structure of the TV-GAN from [29] 12
2.6 favtGAN training procedure from [38] 13
2.7 The framework of ThermalGAN from [39] 14
2.8 Procedure for CycleGAN training, from [1] 15
2.9 Cyclic-GANs' discriminator architecture 16
2.10 Typical cyclic-GAN generator model architecture 17
2.11 The framework for thermal to visible facial translation from [30] . . 18
2.12 The framework of unsupervised-image-generation enhanced adaptive thermal object detector from [40] 18
2.13 The framework of PCSGAN for the thermal to visible facial translation from [41] 19
2.14 The framework of styleGAN2 for thermal face generation from [42] . 20
2.15 The framework for thermal image generation from [43] 21

3.1 An overview of the one-sided translation using patchwise contrastive learning framework. 27
3.2 The Generator Architecture for proposed CUTV2T 29
3.3 The patchGAN discriminator architecture 30
3.4 Comparison between the Least Squares Decision Boundary and the Sigmoid Decision Boundary for Updating the Generator derived from[51]. 32
3.5 Workflow of Patch-wise Contrastive loss 32

4.1 Example images taken from the Carl database [2] 36
4.2 Images which are taken from the Tufts Face Database an example [3] 38
4.3 The (SSIM) Structural Similarity Measurement system's layout and flow from [59] 40
4.4 Architecture of the RetinaFace model [63] 44
4.5 Translated results for CUTV2T(even). 46
4.6 The translated results for CUTV2T(odd) 47
4.7 The translated results for Ext only. 47
4.8 The translated results for Last. 48
4.9 The translated results for no id. 48

4.10 Input images from the Carl database[2] 52
4.11 Results of CUTV2T model on Carl database[2]. 52
4.12 Results of CycleGAN [1] model on Carl database[1]. 53
4.13 Results of pix2pix [4] model on Carl database.[2] 53
4.14 Input/source images from Tuft database [3]. 53
4.15 The proposed CUTV2T results on the Tuft facial database [3]. . . 54
4.16 Results from the CycleGAN [1] using the Tuft face database[3] . . 54
4.17 Results from the Pix2pix [4] using the Tuft face database [3]. . . . 54

List of Tables

2.1 Comparitive analysis of the literature review 23

4.1 The quantitative results with different parameter settings 45
4.2 Quantitative Comparison with CycleGAN [1] and pix2pix [4] on Carl database [2]. 52
4.3 Quantitative Comparison with CycleGAN [1] and pix2pix [4] on Tuft database [3] 52

Abbreviations

ADA	Adaptive Discriminator Augmentation
CG	Computer Graphics
CV	Computer Vision
CCA	Canonical Correlation Analysis
cGANs	Conditional GANs
CUT	Contrastive unpaired translation network
CycleGAN	Cycle-Consistent Adversarial Network
DL	Deep learning
DCGAN	Deep Convolutional Generative Adversarial Network
enc	Encoder
FID	Frechet inception distance
FR	Face Recognition
favtGAN	facial-visible-thermal-GAN
GANs	Generative Adversarial Networks
GNR's	GAN's N Roses
IP	Image processing
I2I	image to image transformation/translation
PSNR	Peak signal to noise ratio
PR	Pattern Recognition
PCSGAN	Perceptual Cyclic-Synthesized Generative Adversarial Networks
T2V	Visible to thermal Transformation of facial images
UQI	Universal quality index
VAEs	Variational autoencoders
V2T	Thermal to Visible Transformation of facial images

Symbols

D_Y	Discriminator which aims to distinguish between real authentic data and generated authentic data
D_X	Discriminator which aims to distinguish between real
E	Expectation
F	Generator model mapping from domain Y to domain X
G	Generator model mapping from domain X to domain Y
L_{GAN}	Adversarial loss
L_{CYC}	Cycle consistency loss
$L_{identity}$	Identity loss
	synthetic data and generated synthetic data
$P_X(x)$	Statistical distribution of the stochastic variable x
$P_Y(y)$	Statistical distribution of the stochastic variable y
$Patch_{NCE}$	Patch-wise noise contrastive estimation loss
$\|\|\|_1$	$\mathrm{L}_1 - norm$

Chapter 1

Introduction

1.1 Motivations

Within the last three decades, Visible face recognition has become among the main fascinating research topics in machine learning, computer vision, and, more recently, deep learning. In deep learning, these visible face recognition (FR) algorithms typically use databases containing tens of thousands or even more visual images. This visible facial recognition technology has progressed to the point that it can be employed in actual conditions. However, variations in illumination conditions and pose changes are critical aspects that considerably influence the effectiveness of FR algorithms in real-world applications.To solve these issues, researchers studied and demonstrated that employing thermal images in FR applications can circumvent the limitations imposed by the visual spectrum, which include invariance to changing in illuminations robustness to changing in posture [5] [**?**], which are the most important aspects influencing the effectiveness of FR systems, particularly in unconstrained conditions [6].These results can be achieved because of the physical factors of the thermal imaging technologies that operate at long-wave infrared wavelengths (8-12 µm). However, since there aren't many thermal databases in the literature and the ones that are available only have a small number of poor-quality photographs, the data demand is a crucial challenge in FR while using thermal imaging. As a result, it's worth looking into different

approaches to augment the thermal face data with image-to-image transformation.

Moreover, surveillance systems with real-time face recognition are increasingly finding their utility in the modern world security applications. As the accuracy of face recognition systems has improved tremendously due to recent advances in deep learning, the reliability of such surveillance systems have also improved manifolds. However, in night-time surveillance, accurate face recognition cannot be achieved as the thermal images of faces taken in the night are totally different in their characteristics as compared to the visible face images in the gallery. A simple solution to this problem can be to convert the RGB images from gallery to synthetic thermal images so that they have more chance of matching with the corresponding night-time thermal face image taken by the surveillance camera. Therefore accurate prediction of a thermal face image from its visible variant will be vital in such application as the accuracy of matching will be dependent on that. To predict a thermal signature from a visible image is not a straightforward regression task but actually an ambiguous problem as there is no relevancy and information present in a visible image related to its thermal variant. However, the case of face images is different as there is a specific structure of a face with typical temperature variations among different areas, for instance nose has mostly lower temperature and eyes have higher. Therefore prediction of thermal variant of a face image can be more accurate and can have more real-world applicability.

1.2 Purpose

Compared to the work on T2V translation models, there is much less work done on visible to thermal translation models. This is because the prediction accuracy of these systems is mostly below par as sufficient training data to learn the heat signature of different parts of images is not commonly available. This study is motivated by this main limitation, and the goal of the project is to develop a V2T translation model that can be trained even in the absence of proper training data.

As facial images have a typical structure and symmetry based on visible features as well as thermal signatures, it would be reasonable to translate them.

1.3 Background

The field of artificial intelligence has expanded significantly over the past ten years. It aided in the development of new applications by using innovative algorithms, including the prediction of 3D protein structures for amino acid sequences [1] and autonomous driving, etc. Without neural networks and deep learning, none of them have been possible. Deep learning is the branch of machine learning that utilizes various neural network-based architectures to learn a variety of tasks.Today deep learning is gaining a lot of popularity. This popularity is due to the availability of training data in the vast amount. Deep learning has made significant strides in several areas, including voice recognition, image understanding, self-driving cars, natural language processing, and search engine optimization. However, regardless of these developments, there are still many limitations in the deep learning model that prevent its widespread use. Among the limitations of deep learning major limitation is deep learning algorithms require massive datasets for training to obtain desired results. Large tech companies like Google and Microsoft can gather and have a lot of data, while smaller firms with innovative ideas might not be able to do so.

Face recognition (FR) is the standard system of recognizing individuals. FR has been utilized in deep learning using various architectures, essentially centered on convolutional neural networks (CNN).FR currently makes use of architectures like Inception-ResNet [7], vision transformers [8], and algorithms created for effective and reliable face comparison [9]. These architectural designs have achieved amazing outcomes under regulated conditions. However, in practical applications, several factors, like changing lighting conditions and poses, significantly affect such systems' performance [10]. Therefore Developers should use various face images with various poses and lighting to enhance the accuracy of such FR systems.

To make the performance of the visual FR more robust, the researchers studied

and demonstrated that employing thermal images in FR applications can circumvent the limitations imposed by the visual spectrum. Unfortunately in the literature until now, no extensive thermal face database is available to train any deep learning-based face recognition.Therefore, it is important to consider several strategies to augment the thermal facial data synthetically.

The I2I transformation aims to translate the input photo of the source domain to the target domain while preserving its inherent source material or content and transferring the target style. I2I transformation that transfers a photo from the input domain image to the target domain image could encompass various challenges in IP, CG, and CV. Particularly I2I transformation could be applied to a wide range of applications, including synthesis of semantic images [11], [12] ,image segmentation [13],[14], [15], style transfer [1],[16],[17] and image inpainting [18], [19], [20] etc.

As discussed in section 1.1 this research is motivated by the fact that visible to thermal transformation of facial images is still a challenging problem especially when there is a lack of training data.

1.4 Structure of the Thesis

The structure of the thesis is as follows:

- The chapter1 covers the motivations, purpose, and background. The structure of the thesis is also covered in this chapter.
- Chapter 2 describes the literature review. It includes information regarding GAN's-based and some non-GAN approaches. Additionally, the chapter also covers the Gaps, Problem statement, and research contributions.
- Chapter 3 thoroughly explains the methodology used. The chapter includes information about Background, loss functions, and architectures used.
- Chapter 4 describes the datasets used, which include the details of the Carl database and the Tufts Face Database.
- Chapter 5 provides details of the metrics used to evaluate the effectiveness of the model, including the FID, UQI, PSNR, and SSIM.

- Chapter 6 thoroughly provides the details of the experimentations done. It includes training details, a discussion on results, the hardware used, and the quantitative and qualitative comparison.
- The final chapter is about the conclusion, future work, and challenges.

1.5 Summary

This chapter introduced our task. It contained details about motivations, purpose, and background. The background section of this chapter briefly discussed the I2I transformation terminology and its applications in different fields such as PR and CV, etc.

Chapter 2

Literature Review

2.1 Introduction

An "image-to-image translation" is the term used to describe a large set of tasks that involve learning a mapping between images in one distribution and images in another, where the distributions and desired features of the function differ depending on the underlying task involved. The examples include synthesizing images based on segmentation, image improvement, super-resolution of images, the transformation of an object, colorization of the image, style transfer, denoising of images etc. These different tasks can be merged into a single problem with the advent of neural networks : using pairs of examples from both domains, educate a convolutional neural network to map input images to output images.

The goal of I2I transformation is to map between different domain images. Therefore, representing these mappings to achieve the desired outcomes is directly linked to generative models.The generative model[21],[22],[23], suggests that the data are generated via specific distributions. Suppose the distribution is Normal or Gaussian distribution. In that case, two parameters need to be calculated to estimate the distribution, i.e., a mean and standard deviation.if the distribution is a Non-Gaussian or Non-parametric (Each sample contributes to the distributions independently), these underlying distributions are estimated with specific techniques.Today Deep generative models have demonstrated tremendous progress in

estimating missing data [24], formulating predictions [25], compressing datasets [26], and producing unseen data.

In an I2I transformation, the generative model simulates output domain distribution by generating convincing "fake" images so that transformed images appear to be drawn from the distributions of the target domain. Particularly for I2I transformation task, two of the most often utilized and effective generative models are the VAEs and GANs. Although the methodologies of both models are distinct, both models strive to develop a replica x = g(z) to generate the required samples x from a latent space z.

2.1.1 Variational Autoencoder (VAEs)

In deep learning, Variational autoencoders (VAEs) are a technique for learning good latent representations.The VAEs [27],[28], contain two networks, an encoder, and a decoder.

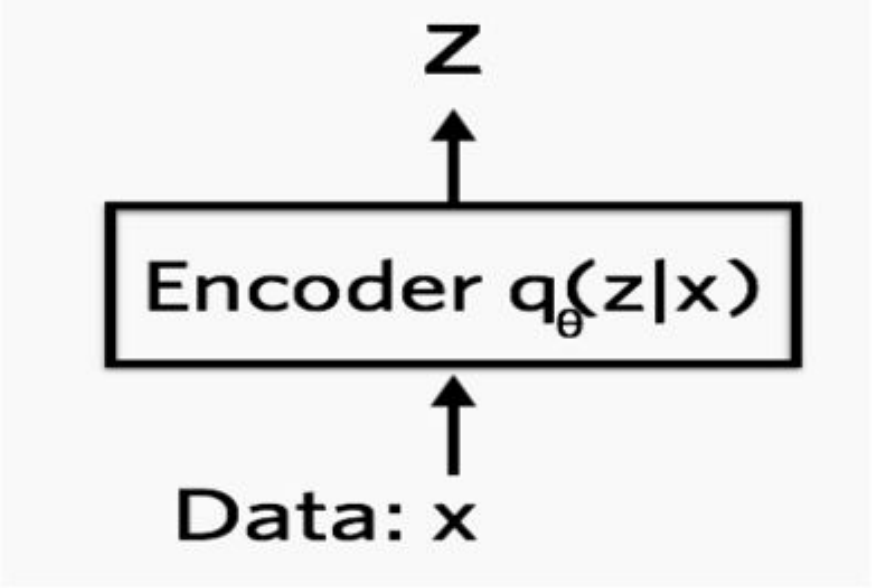

FIGURE 2.1: Encoder compresses data in the Latent Space (Z)

As shown in Figure 2.1,the encoder takes a data point x as input, generates a concealed representation z as output, and has weights and biases. For instance, suppose x is an image of a handwritten number with a resolution of 28 by 28

pixels (i.e.,784 dimensional).). If this input x which has 784 dimensions, is passed through the encoder, then the encoder learns its hidden representation z, and the dimension of the latent representation is much smaller than its input dimension, i.e., 784. So encoders are required to master effective data compression for this lower-dimensional space, commonly referred to as a "bottleneck". The encoder is represented as $q_\phi(z|x)$. Commonly the encoder output distribution is assumed as the Gaussian distribution.

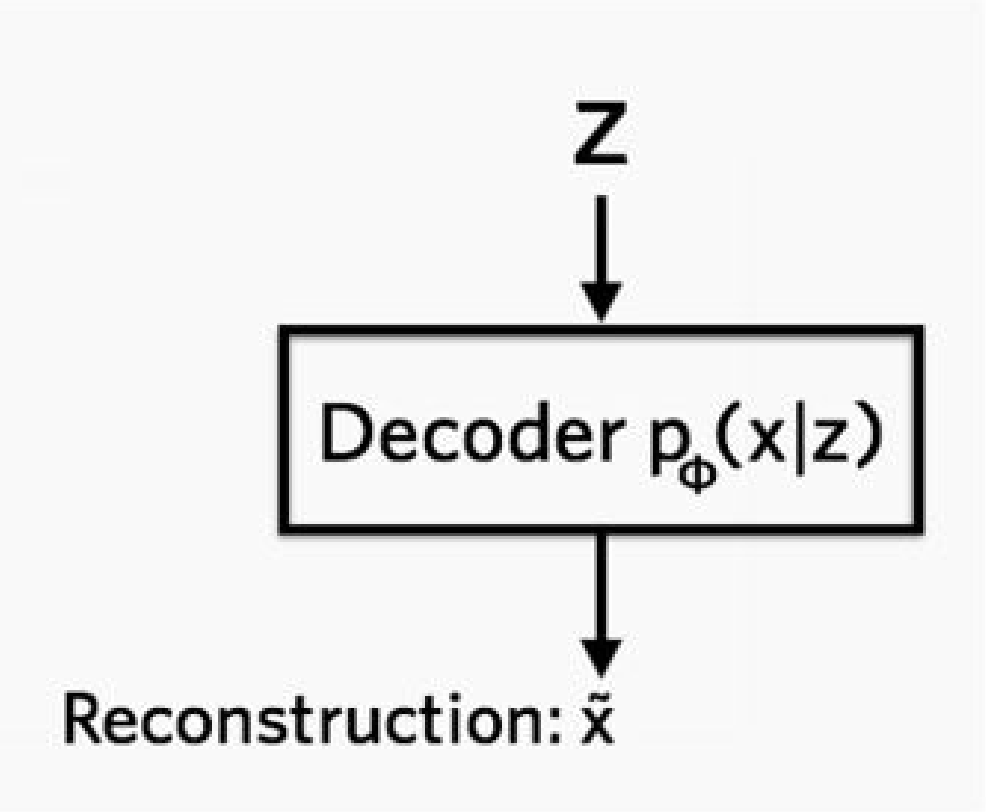

FIGURE 2.2: The Decoder Reconstruct the Data Given the Hidden Representation

In a variational autoencoder, the other network is a decoder network. As shown in Figure 2.2, the decoder takes the latent representation z as input and outputs the parameters for the data probability distribution and also has weights and biases.

The purpose of the decoder is to reconstruct the input image from the latent representation z, which contains the same dimensions as the input image. Following

the example of the handwritten digits, let's assume that the photographs are in the format of black and white, and every pixel is represented by a 0 or 1. To estimate each pixel, the decoder utilizes the Bernoulli distribution representing the single-pixel probability distribution. So for each pixel, the decoder returns a Bernoulli parameter. If the image contains 784 dimensions, the decoder returns 784 Bernoulli parameters. This way, the decoder decodes the real number within the z vector into 784 real numbers ranging from 0 to 1.

As the data summary or the compressed representation of the input image is available for the decoder network, the information of the original image is not ideally recovered. What percentage of the data is lost? The reconstruction $log_p(x|z)$, with nats as units, is used to determine this. And this measure indicates how well the decoder had learned to recreate the input picture x from its hidden representations z.

Next, we discuss various possible techniques that have achieved the image synthesis goal from visible to thermal spectrum or from thermal to visible spectrum. It's worth noting that, while attempts have been made to synthesize images from thermal to visible for facial photos, there has been very little work done to synthesize images from visible to thermal spectrum. This is mainly due to the lack of relevant training data and the inability of existing or previous models to capture the heat signatures of everyday objects. The following subsections summarize numerous works that sought to execute synthesis from one domain to another.

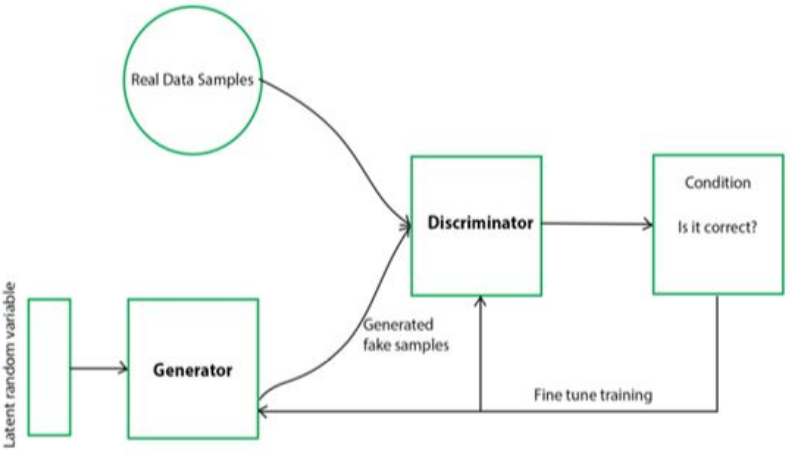

FIGURE 2.3: Generative Adversarial Network's workflow

2.2 GANs Based Synthesis

In the recent advancement in deep learning, a lot of research is centered on GANs to generate visual image data not exclusively from thermal image data [29], [30] but also from polarimetric [31],[32],and near-infrared image data [33],[34] GANs were firstly introduced by I. Goodfellow in [35],which can try to create from any input data distribution through a com¬petition between a generator and a discriminator neural network.

GANs was a significant step forward in image synthesis, as they were the first models to create photorealistic images from random noise. As shown in Figure 2.3, a GAN comprises an image-outputting generator network and a discriminator network that has been educated to discriminate among the generator's outputs and authentic images from its distribution of interest.

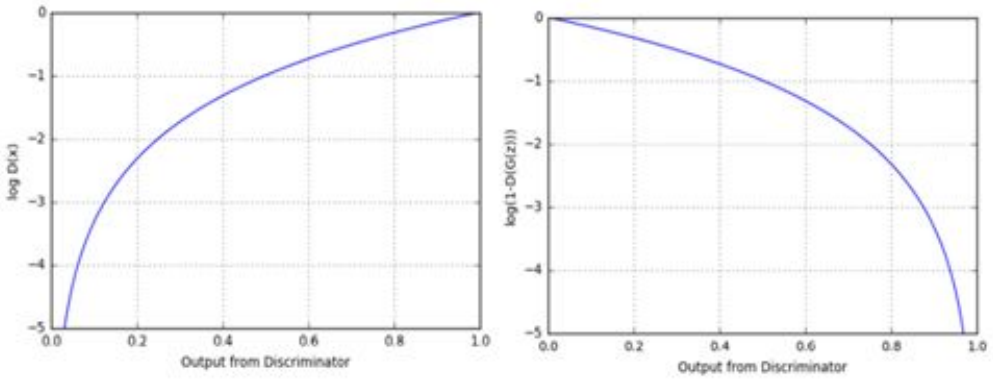

FIGURE 2.4: GAN loss function terms, from [35]

Simultaneously, the generator is trained to produce pictures that prevent the discriminator from making this distinction. This adversarial strategy, when properly trained, can aid the generator in producing realistic outputs, since a skilled discriminator can correctly categorize images containing artifacts as synthetic. The minimax loss provided is used to define the optimization problem. Examining the two log functions utilized in the GAN loss function explains the GAN loss function in equation (2.1). Figure 2.4 shows the log terms used to demonstrate the goal function. The aim of the discriminator is to maximize the values of the two functions in Figure 2.4.When the discriminator's input is an authentic image, the

discriminator's aim is described by the left plot in Figure 2.4.The discriminator's goal is to maxi¬mize log D(x), where x is an actual image, and an output of 1 refers to the prediction that a photo is accurate. A prediction that an image is false is shown by a 0 out¬put. As a result, the discriminator's goal is to maximize log(1-D(G(z)), where z is the generator's noise input which is described by the right plot in 2.4. Similarly, the generator's goal is for the discriminator to output a 1 due to bogus image input. As a result, the generator's goal is to reduce log(1-D(G(z)), as shown in Figure 2.4.The model should approach Nash equilibrium in an ideal environment in which it converges.

The generator makes bogus images from random noise fed into the dis¬criminator, while actual photos are also supplied to the discriminator. The generator aims to create fictitious images that the discriminator recognizes as authentic. While the discriminator's objective is to accurately classify both fake and real images.

Furthermore, GANs feature a one-of-a-kind Nash equilibrium where the gen-era¬tor samples precisely from the target distribution. The discriminator might label some areas of the target distribution as synthetic if the generator oversamples them. Mode collapse is another famously challenging difficulty to over¬come in reality, despite several works that have devised ways to resist the phenomenon. We discussed GANs in the context of image synthesis, the generator's input may be any distribution, making it appropriate for image-to-image translation challenges. The outputs of the U-Net mentioned previously become photorealistic when adversarial training is integrated into the loss function; pix2pix or a conditional GAN are terms used to describe the gener¬ated model.

Several variants of fundamental GANs have been presented, which can apply to synthesis tasks. Examples are Boundary Equilibrium Generative Adversarial Networks (BEGAN)[36], and Deep Convolutional Generative Adversarial Networks (DCGAN) [37].BEGAN [36],introduced an equilibrium factor that regulates model training by balancing the discriminator and generator.While in DCGAN [37] , a convolution neural network (CNN) is incorporated into the generator and the discriminator. These variants of GANs model significantly enhanced the training stability while not enhancing the generated image quality.

The introduction of conditional generative adversarial networks (cGANs) [4],was

a huge breakthrough in the research related to image-to-image translation. In applica¬tions as diverse as sketch-to image conversion, image colorization, image inpainting, and style transfer, cGANs provide state-of-the-art outcomes. Recent studies have shown some exciting image-to-image translation applications of cGANs, in¬cluding generating face images of younger or older ages, creating zebra images from horse images, etc.

2.2.1 GANs for Visible to Thermal Image Translation

The visible-to-thermal image transformation/translation remains a difficult challenge, and cGANs have had less success in this arena. Because the cGAN requires paired data that are tightly linked to one another, and the shapes or objects involved in the images have a clear specific structure.

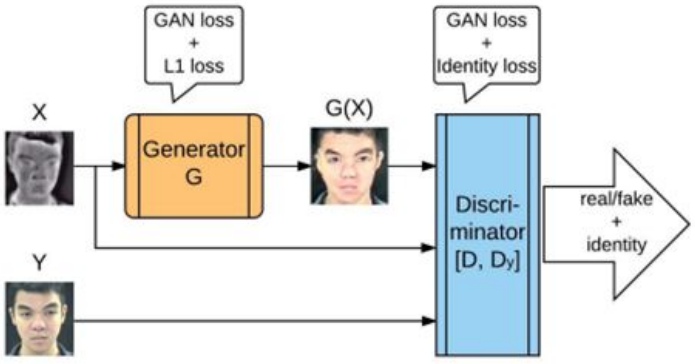

FIGURE 2.5: The structure of the TV-GAN from [29]

Two variants of GANs are largely used for the translation task. The first one is Image-to-Image Translation with Conditional Adversarial Nets (Pix2Pix) [4] , and the other is Cycle-Consistent Adversarial Networks (CycleGAN) [1]. Pix2Pix [4], can achieve decent results, but the limitation of Pix2Pix [4], is that it requires paired data as input, and paired data is rarely available in case thermal and visible spectrum domains to precisely pick the thermal signatures of different objects. CycleGAN [1], on the other hand, is designed to work with unpaired training

images, making it more suitable for applications where training data is scarce. For other applications both Pix2Pix [4] and CycleGAN [1] demonstrated their efficacy in generating high-resolution pictures but at the cost of computational complexity due to complex topologies and the incorporation of post-processing techniques. More importantly, training such large models is computationally costly and requires large databases that are either unavailable or available with strict copyrights, which makes them impossible to use to achieve satisfactory results.

2.2.1.1 Pix2Pix Based Approaches

Zhang et al.[29], considered synthesizing colored faces from thermal images with various head poses and occlusion with eyeglasses.This work combined Conditional GANs, which were inspired from the Pix2Pix model [4], with a closed-set face recognition loss to preserve the face identity information as shown in Figure 2.5.The evaluation of cross-spectrum face recognition is then performed, utilizing the MatConvNet VGG-based model previously trained. The results showed a performance improvement compared to the claimed performance of the Pix2Pix system [4].

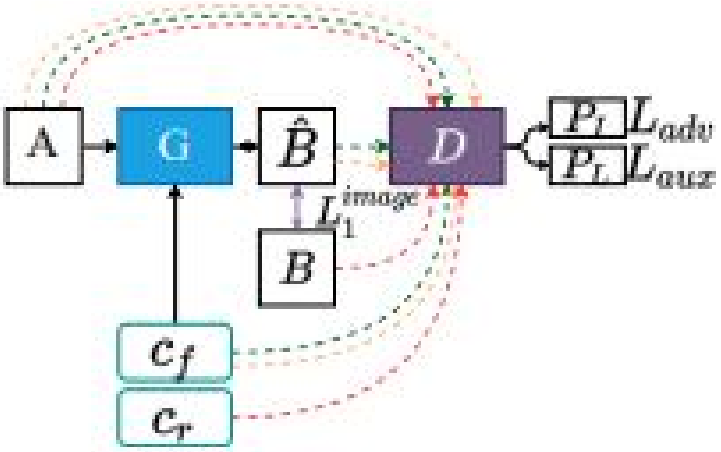

FIGURE 2.6: favtGAN training procedure from [38]

Ordun et al.[38] proposed a favtGAN, that generates thermal face images from visual images by combining the pix2pix image translation model with an auxiliary sensor label prediction network. Firstly face and cityscape databases were integrated. This integrated database was taken through a similar sensor to bootstrap the training and transfer learning process. The result showed a performance improvement in terms of SSIM and PSNR scores of generated thermal faces with combined datasets compared to training on a single dataset alone. Figure 2.6 shows the block diagram of favtGAN [38] model.

A significant study was reported by V. V. Kniaz et al, in which the authors proposed a dedicated architecture for colored to thermal image translation, called ThermalGAN [39]. ThermalGAN[39] is a pipeline in which the first generator, BicycleGAN [1], [19] with some minor adjustments to U-NET, produces multi-modal pictures of segmented masks, and the second generator is based on Pix2Pix model [4] which is shown in the Figure 2.7.ThermalGAN is trained and evaluated on a single dataset from a single sensor, and it is designed for surveillance applications, not only for faces, but for the complete body.

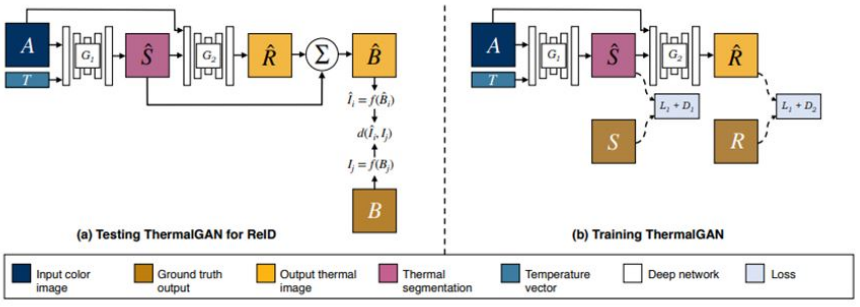

FIGURE 2.7: The framework of ThermalGAN from [39]

2.2.1.2 CycleGAN

As we are interested in models which can be trained with unpaired data, CycleGAN is one the main candidate model to be used in visible to thermal image translation. Therefore, here we present an overview of CycleGAN.

Zhu et al. [1],[19] proposed the cycleGAN and displayed remarkable image-to-image transformation/translation performance, building on the success of Generative Adversarial Networks. Two generators and two discriminators make up their model. In place of a noise vector, the generators accept the image as input and produce the image that keeps the identity of the input image while only altering the domain (e.g, tuning a horse to a zebra while keeping another generator). This output image is then compared to the input image to ensure that only the domain changes and not the identity of the image. The cycle consistency, also known as reconstructing loss, is important because it motivates a model to keep the identity of the image intact by avoiding changing data that is unnecessary to change. The model is instructed by the reconstructing loss to change only an image and leave everything else unchanged.

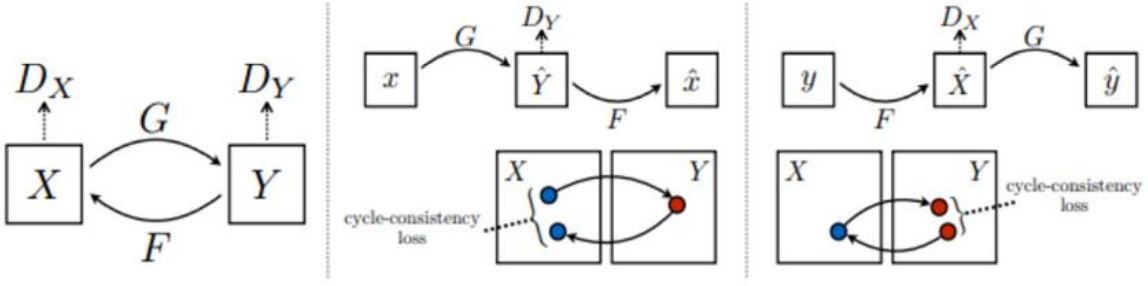

FIGURE 2.8: Procedure for CycleGAN training, from [1]

In order to provide information to a generator during the expected domain shift, the discriminators categorise the image as real or fake. The input is rebuilt from the generated image using domain and keep the rest unchanged. When converting a horse image to a zebra image, for example, the model's goal is to just convert the horse to a zebra. The lack of cycle consistency pushes the model to maintain the image's trees and other background information. Figure 2.8 demonstrates the process. G is an image translator that converts images from domain X to domain Y, A generator named F converts images from the domain Y to the domain X, In domain Y D_Y serves as a discriminator for authentic and fake images, and In domain X D_X serves as a discriminator between actual and fake images.

The goal of the CycleGAN [1] model extends the traditional GAN goal by taking the reconstructing loss into account. An L_1 loss between both actual and regenerated inputs determines the reconstructing loss. The goal may be seen in:

$$min_{G,F}max_{D_X,D_y}L(G,F,D_X,D_Y) = L_{GANs}(G,D_Y,X,)+L_{GAN}(F,D_X,Y,X)+\lambda_{LCYC}(G,F) \tag{2.1}$$

The L_{GANs} target is (2.1), and the cyclic loss is L_{cyc}, as illustrated in (2.2)

$$L_{CYC}(G,F) = E_{x\sim Pdata}[\|F(G(x)-x\|1] + E_{y\sim Pdata}[\|G(F(y))-y\|1] \tag{2.2}$$

In a CycleGAN, the discriminator architecture is quite similar to that of a regular CNN for binary classification. The network takes an image as input and returns a single prediction between 0 and 1. A forecast of 0 means there's a 100 percent chance the input image is fake, while a prediction of 1 means there's a 100 percent chance the input image is real. Convolutional layers are used in the network, which is proceeded by the batch normalizations as well as a ReLU activation function. Figure 2.9 shows the discriminator architecture.

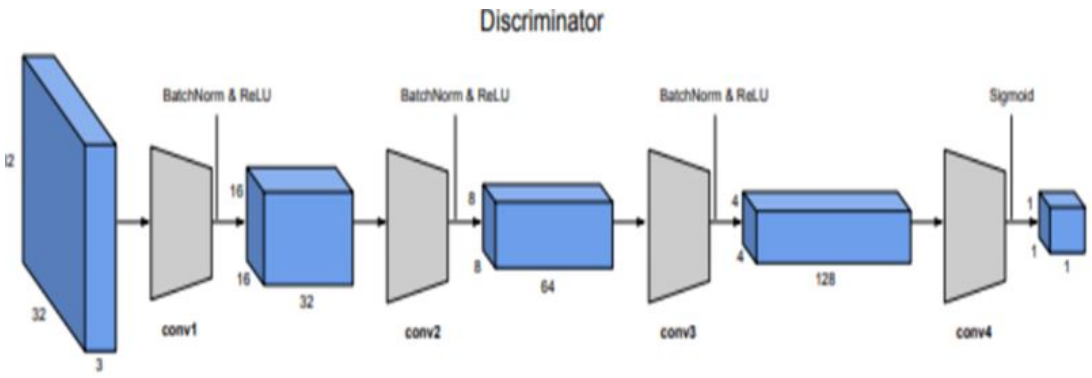

FIGURE 2.9: Cyclic-GANs' discriminator architecture

Following the first layer and excluding the last layer, as indicated in Figure 2.9, every convolutional layer downsamples the spatial dimensions by a factor of two. At the same time, the number of channels rises by a factor of two. After the last convolutional layer, a sigmoid activation is utilized to reduce the prediction to 0

and 1.

The generator architecture is special as its input and output are both images. As a result, given a CycleGAN [1], generator design will have numerous components. First, a convolutional layer sequence includes a ReLU activation function and batch normalization layer comparable to a discriminator model architecture. A residual block is frequently seen in the second part network. Finally, the network's final layer consists of de-convolutional layers that up-sample the sample to image size. Figure 2.10 shows a typical CycleGAN Generator architecture.

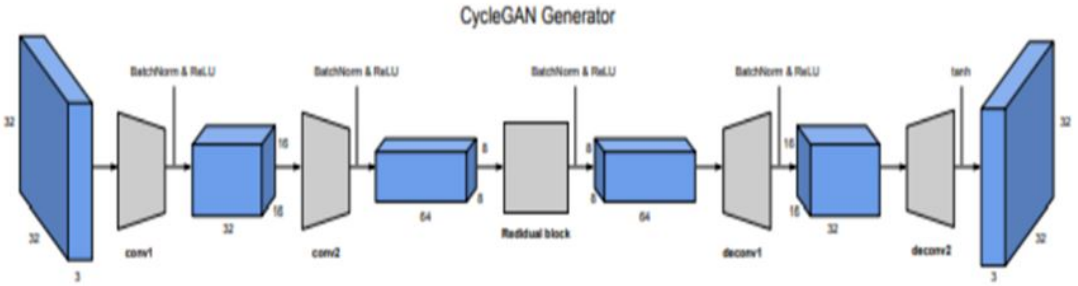

FIGURE 2.10: Typical cyclic-GAN generator model architecture

The three sections of the CycleGAN generator are depicted in the generator architecture shown in Figure 2.10. The first section's convolution layers all downsample the spatial dimensions by a factor of two, similar to how the dis¬criminator does. Starting with the second layer, each convolutional layer doubles the number of channels. The third section is exactly the opposite, i.e.; each deconvolution layer doubles the spatial dimensions while halving the number of channels to restore the original image dimensions.

Although CycleGAN [1] has demonstrated acceptable results in cross-domain image to image translation, the problems which were studied were relatively simpler as compared to our case of visible to thermal translation. CycleGAN [1] has not been evaluated against more challenging problems like the one at hand.

2.2.1.3 Variants of CycleGANs for Image Translation

Wang et al.[30] proposed a framework that is derived from the CycleGAN [1] model included a facial landmark detector loss that depicts face identity preserving

features for converting a thermal facial image to a visual facial image as shown in the Figure 2.11.

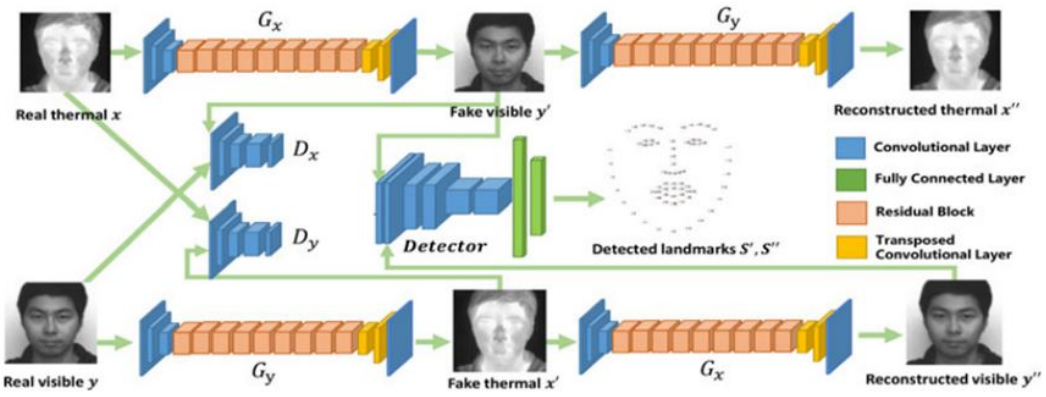

FIGURE 2.11: The framework for thermal to visible facial translation from [30]

This system was evaluated using a FaceNet model pre-trained on publicly accessible visible datasets and improved cross-spectrum face recognition performance as compared to the original CycleGAN [1] model.

Li et al.[40] constructed thermal pedestrian landscapes from visible images as a data augmentation technique for a downstream object detection job.

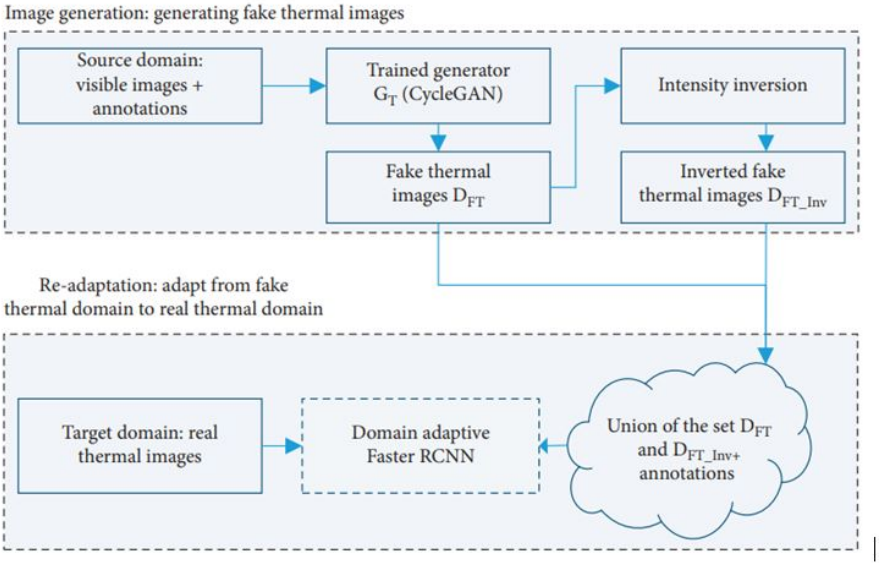

FIGURE 2.12: The framework of unsupervised-image-generation enhanced adaptive thermal object detector from [40]

For the landscape generation, they use CycleGAN [1], followed by an intensity inversion transformation [40], as shown in Figure 2.12.

A significant study was reported by Babu Dubey, in which the authors proposed a dedicated architecture for Thermal and NIR to Visible Image transformation/-translation, called PCSGAN [41].

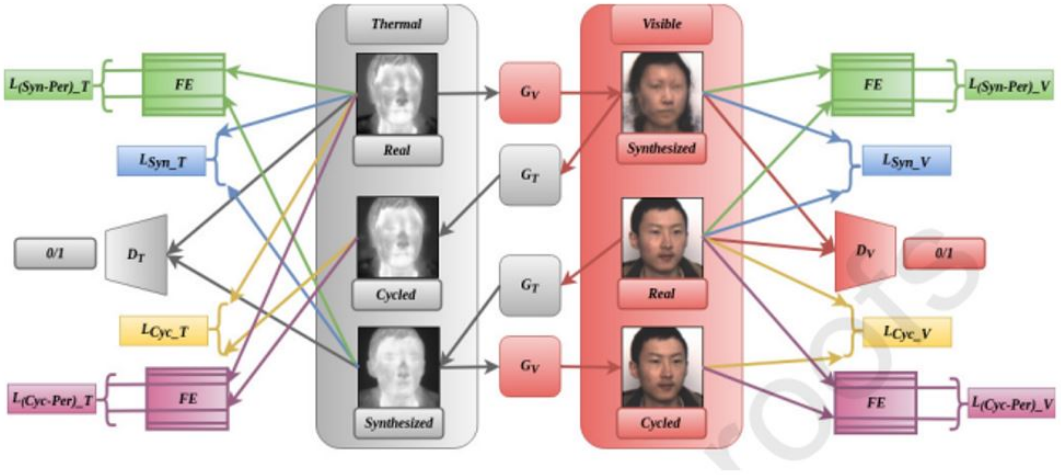

FIGURE 2.13: The framework of PCSGAN for the thermal to visible facial translation from [41]

Similar to DualGAN [17], and CycleGAN [1], the PCSGAN [41], employs two generators and two discriminator networks as shown in Figure 2.13.PCSGAN [41], converts the NIR infrared or thermal face image to a visible face image by combining three loss functions, i.e., perceptual (feature-based) , pixel-wise and adversarial loss. Also the PCSGAN [41], model was evaluated both qualitatively and quantitatively on the WHU-IIP face and RGB-NIR scene datasets and showed improved performance in terms of SSIM, PSNR, and MSE as compared to state-of-the-art image translation models such as Pix2Pix [4], DualGAN [17], and CycleGAN [1].

2.2.1.4 Other GAN Architectures for Thermal Face Generation

StyleGAN2 [42], which is an extension of StyleGAN, that is used to generate different augmented variants of a thermal image.Although this is not strictly cross-domain I2I transformation but the authors proved that even a limited database

can be used and sizeable number of the replica with different pose variations can be generated to train larger networks.StyleGAN2 allows the synthesis of excellent-quality and high-resolution images. However, training this kind of network with fewer data can be difficult.When working with a small data set, the discriminator adjusts immediately to the training instances ,generating overfitting.To solve this issue, StyleGAN2 [42], was improved by utilizing adaptive discriminator augmentation, known as StyleGAN2-ADA.This method entails requesting the generator to provide samples that cannot be separated from the training set when seen via all of these distorted glasses and testing the discriminator with just magnified images to prevent the generator from learning the augmented distribution (leak).The authors look at the augmentation probability p, which gives us control over what the discriminator sees. This is how StyleGAN2-ADA allows you to use the augmentation probability to transform non-invertible data augmentation into invertible transformations. The authors show that GANs via Style-GAN2 are used to generate synthesized thermal images. Along with the StyleGAN-ADA, various StyleGAN2 versions are used. In addition to training StyleGAN2 and its many versions, available thermal databases were also utilized for training a thermal face detector based on YOLOv3. During the training process, various metrics were also evaluated.Figure 2.14, is shown below,represents this methodology.

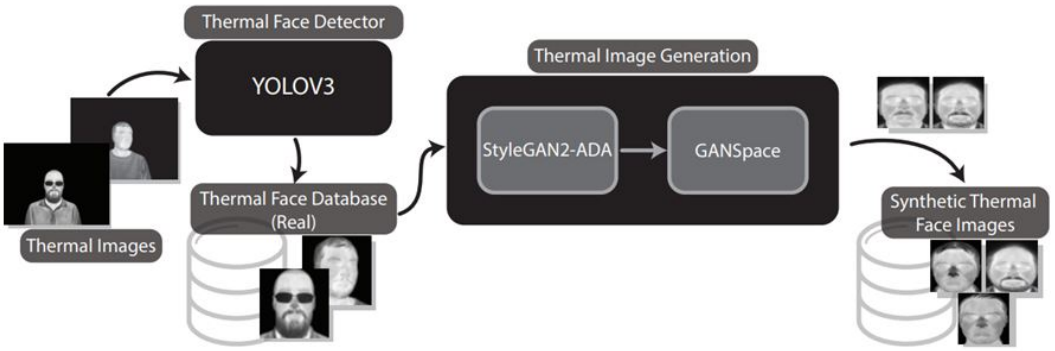

FIGURE 2.14: The framework of styleGAN2 for thermal face generation from [42]

Based on the various deep learning models Pavez et al.[43] proposed a method to create thermal images with six various aspects (smile, vocal, frown, glasses, normal, and rotation) for robust face recognition. Firstly style clip is utilized for

the manipulation of the input visible image latent space to add desired attributes to the visual face image. Then GANs N' Roses (GNR) model is utilized which uses the style and content maps to create a thermal face image from the visual image utilizing the adversarial approach as shown in Figure 2.15.

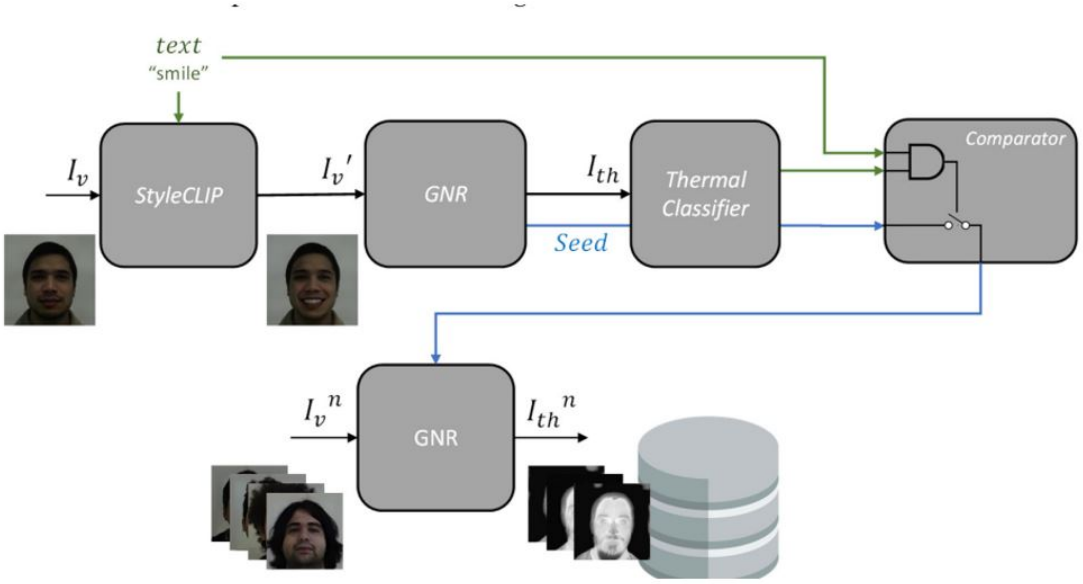

FIGURE 2.15: The framework for thermal image generation from [43]

2.2.2 Image Translation Network Architectures other than GANs

2.2.2.1 ThermelNet

In another work, synthesis of the thermal image is achieved by revising the Residual SqueezeNet known as ThermalNet [44].The ThermelNet [44], has fourteen layers of convolution, deconvolution, and re-modules. To maintain the spatial resolution of the input source image firstly two more deconvolution layers have been added to the architecture of the residual SqueezeNet. Afterward, the global avgpool layer was detached from the residual squeezeNet architecture.Once the thermal images had been created a trained VGG-16 network was used for post-processing. During the output thermal image synthesis, the network serves as a similarity benchmark to ensure that the image matches the ground truth. The Torch7 library was used to run the network. The method of employing a deep convolutional network

to imitate a style is based on an iterative selection of the required image, with the network acting as a measure of the style resemblance. Firstly a Gaussian white noise is utilized to initialize the image.Then Gradient descent is utilized to generate a new image with the matching style, After that, the original image is altered until it produces similar results in a particular the same network layer as the initial image.Here the purpose of the post-processing step was to reduce the quadratic loss error between the generated picture and its target image.

2.2.2.2 Synthesis Based on Local Linearity

Li et al.[45] were the first to investigate the change in the spectrum from thermal to visual.Their work developed a framework that exploited the Local Linearity of the image's spatial domain and its manifold. In addition, the Markov Random Fields were also utilized in their work to arrange the patches of the images in order to enhance the estimated visible face images.

2.2.2.3 Synthesis of Visible Faces Using Local Linear Regression

To perform one-to-one mapping among visual faces and thermal faces, Dou et al. [46] used Canonical Correlation Analysis (CCA) to extract features. Then to understand the association between the two feature spaces, where the visual features are predicted from the correlated thermal features Locally linear regression is utilized. Finally, the visible face is reconstructed using locally linear embedding from the transformed thermal features.

2.2.2.4 Cascaded Refinement Network (CRN)

In order to synthesize images from visible to thermal spectrums, the cascaded refinement network (CRN) [47],[**?**] method is used. The first CRN was used to create photographic images from semantic layouts. The proposed architecture grows smoothly to high-quality images, producing photo¬realistic images with a resolution of 2 megapixels from 2D semantic label maps. The difficulty comes

TABLE 2.1: Comparitive analysis of the literature review

No	Ref.	Approach	Major Contributions	Limitations
1	(Thermal2Visible GAN) by Zhang et al.[29]. (2018)	• Used the pix2pix translation model with closed set face recognition loss.	• Presented TVGAN that can keep track of enough identity data. • Without making any changes, the strategy enhances the current VLD face recognition system.	• Required paired data. • Does not guarantee the accurate translation of other facial characteristics, such as race and age. • The technique is for the thermal to visible face translation.
2	FavtGAN by Ordun et al.[38]. (2021)	• Used the pix2pix model in conjunction to auxiliary label prediction network.	• Improve thermal image synthesis by bootstrapping image translation with additional data from other domains. • Visible to Thermal translation model.	• Required paired data • Required data that have a similar optical property.
3	ThermalGAN Kniaz et al.[39] (2018)	• Two stacked GANs i.e BicycleGAN and unimodal pix2pix model.	• ThermalGAN[39] framework. • A sizable Thermal-World multispectral dataset.	• Computationally expensive. • Required paired data in large amount. • Not specific for faces.
4	PCSGAN Babu and Dubey[41] (2020)	• Along with perceptual losses and adversarial losses pixel-wise losses is utilized.	• A new technique for converting thermal/NIR face images to visible ones.	• Computationally intensive. • Required paired data. • Not for the visible to thermal transformation.
5	Transformation of Thermal face image to Visible Face Image Wang et al.[30](2018)	• Generative adversarial network with the detector network.	• A technique for converting thermal facial photos into visual ones. • Approach beat his previous approaches in performance.	• Required paired data. • The approach is for thermal to visible face transformation.

from creating detailed images from simple semantic label mappings. In order to basically make it scale and rotation invariant, training is done using contextual loss. The CRN is a specific kind of CNN with linked refining modules. The first module considers the space with the lowest resolution, in this case, 4x4. The above resolution is then doubled in each consecutive module till it achieves 128x128 in the current scenario, which is the desired image resolution. The style loss is then determined between the actual and generated thermal picture. These losses are then calculated only at the embedding level, and VGG-19 and CRN are used to extract them.VGG-19 is a post-processing tool in this case.

2.3 Gap Analysis

The above section details various studies which have reported work on visible to thermal image translation. Following gaps have been identified in the literature which should be worked upon.

- Thermal to visible face image conversion has been studied in various studies but there are very limited studies addressing visible to thermal image translation problem and those also report translation of rather simple images.
- Two main GAN based architectures proposed for image translation are Pix2Pix and CycleGAN. Pix2Pix can be used for paired training images and CycleGAN can also be used for unpaired training data. However, both these architectures are very complicated needing a lot of data to train which is already very limited in the case of thermal images due to obvious reasons.
- The Visible to Thermal Transformation of facial images using unpaired data is not explored so far to the best of our knowledge.

2.4 Problem Statement

Transformation of visible to thermal images is a challenging task because of the scarcity of training data. The data limitation can be overcome if a model is

developed that can be trained on unpaired training images. However, visible to thermal translation using unpaired training images specially for face images is still an open research problem.

2.5 Research Contributions

In an attempt to bridge the gap discussed in the prior section, the following cardinal and novel contributions have been made in this research thesis:

- A new visible to thermal translation network based on CUT has been proposed that is trained on unpaired facial images.
- The comparison has also been made with other well known translation model including Pix2pix and CycleGAN and shown that the CUTV2T has shown superior performance.

2.6 Summary

This chapter provided the literature review of the thesis in detail. Models which transform the thermal spectrum image into the visible spectrum or vice versa based on GANs and apart from GANs were discussed. Additionally the Gap Analysis,Problem Statement and the Contribution of the thesis were also discussed in this chapter.

Chapter 3

Methodology

3.1 Background

In machine learning, the main goal is to estimate the data distributions. Various machine learning methodologies are used to achieve the data distribution. Most of the methods involve first defining a model of the data and then estimating its parameters. The parameters of the model distribution are either predicted by a neural network or a neural network directly predicts probabilities of the given input. A softmax classification-based neural network is an example of a such model that predicts the probability distribution of the input data.The one property that the probability density function has to satisfy is that the sum of all elements of probability distribution function must be equal to 1. This property is ensured by the normalizing constant. The normalizing constant is the most difficult thing to tackle. The reason is it translates into an intractable realm and solving high-dimensional integrals is not practical for computations. Micheal Gutmann [48] solved the problem of intractable denominator by comparing two distributions. To compare two entities contrastive learning was used.

The underlying assumption of our approach is to learn a good representation that encodes the shared underlying details among various aspects of high-dimensional signals. In addition, low-level information and local noise are discarded. Approaches that use next-step prediction in high-dimensional modeling make use of

the signal's local smoothness. As the quantity of shared details decreases dramatically when forecasting further into the future so the model needs more global structure to derive.

In high-dimensional data prediction, unimodal losses like mean squared error and cross-entropy loss are not particularly handy. Such models are computationally expensive, and they lose capabilities when it comes to modeling the data x complicated relationships and often neglect the context c. As an example, images can have thousands of bits of details, but Latent variables, like class labels, contain far fewer details (10 bits with 1024 classes). This implies that direct modeling p(x|c) should not be the best way to extract shared information between x and c. Strong conditional generative models that rebuild every aspect of the data are typically required.So in order to predict future information, we encode the target x and context c into the compressed vector representation (using learned non-linear mappings) while maintaining the mutual information of the signals x to the maximum extent. Moreover, By maximizing mutual information among the embedded representations, the underlying latent variables provided by the inputs are also recovered.

3.2 Proposed Methodology

Our goal is to transform the visible face image into a thermal face to produce thermal face images closer to the domain of actual thermal face images.

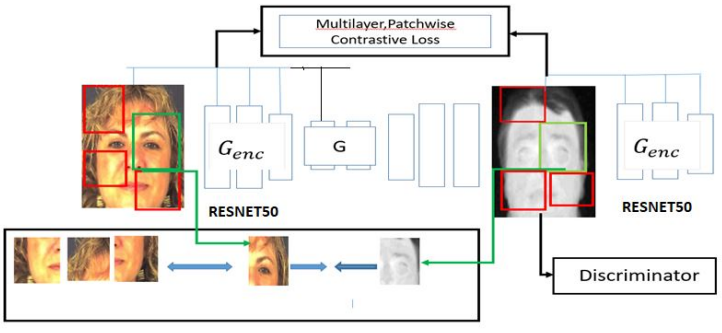

FIGURE 3.1: An overview of the one-sided translation using patchwise contrastive learning framework.

However, transforming visual face images into thermal face images due to the lack of training data is not a trivial task. Therefore a sophisticated translation model is needed.

To perform the translation of facial images, a model called the contrastive unpaired translation network for visible to thermal transformation (CUTV2T) is presented. This model may be trained on unpaired visible and thermal facial images. The CUTV2T selects two related patches (positives) compared to other patches (negatives) in the dataset as shown in Figure 3.1 and then maps both patches to the exact location in the learned features space. In the context of unpaired image-to-image translation, it is shown that the strategy used enables one-sided translation while enhancing the quality and lowering training time.

3.2.1 Architectures

To perform the Visible to the thermal transformation of facial images two networks are utilized, including the generator and the discriminator. The generator generates real-looking images while the discriminator discriminates between authentic and generated images.

3.2.1.1 Generator

Our Generator model consists of an Encoder G_{enc} and decoder G_{dec} which are utilized sequentially to synthesize the output image. The architectural guidelines established by [37] are roughly followed in our generator network. The in-network downsampling and upsampling use stride and fractionally stride convolutions instead of pooling layers. The architecture of [49] is used to create our network body, which consists of nine residual blocks [50]. Batch normalization and ReLU nonlinearities are applied after each non-residual convolutional layer aside from the output layer, which uses a scaled tanh to guarantee that the resulting pixels are within the range [0, 255].7x7 kernels are used in the first and last layers, and

3x3 kernels are used in the rest of the convolutional layers.Figure 3.2, is shown below,represents this architecture.

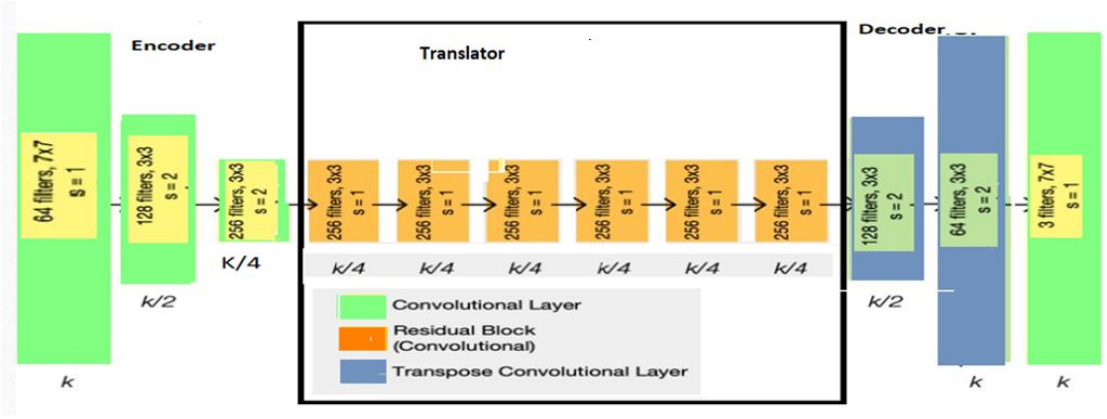

FIGURE 3.2: The Generator Architecture for proposed CUTV2T

The residual connections [50] are used for the network body of the generator. We believe the residual connection makes learning the identity function easier; so this is a crucial ingredient for image transformation architectures because the resulting image should share structure with the source image in most situations. Our network's body comprises many residual blocks, each with two 3 x3 convolutional layers.

3.2.1.2 Discriminator

The basic PatchGAN model is selected as a discriminator network that forecasts whether the 70x70 patch in the input image is genuine or fake. The output of the network can then be calculated by averaging these forecasts. Our discriminator model requires 256x256 images as input and defines an explicit architecture applied to all test problems.

Following is the architecture of the discriminator: C64-C128-C256-C512 as shown in the Figure 3.3 (C64 means convolution layer which contains 64 filters).In the CycleGAN [4] terminology, this is referred to as a 3-layer PatchGAN because, apart from the first hidden layer, the model contains three hidden layers that may

be scaled up or down to produce different-sized PatchGAN models.

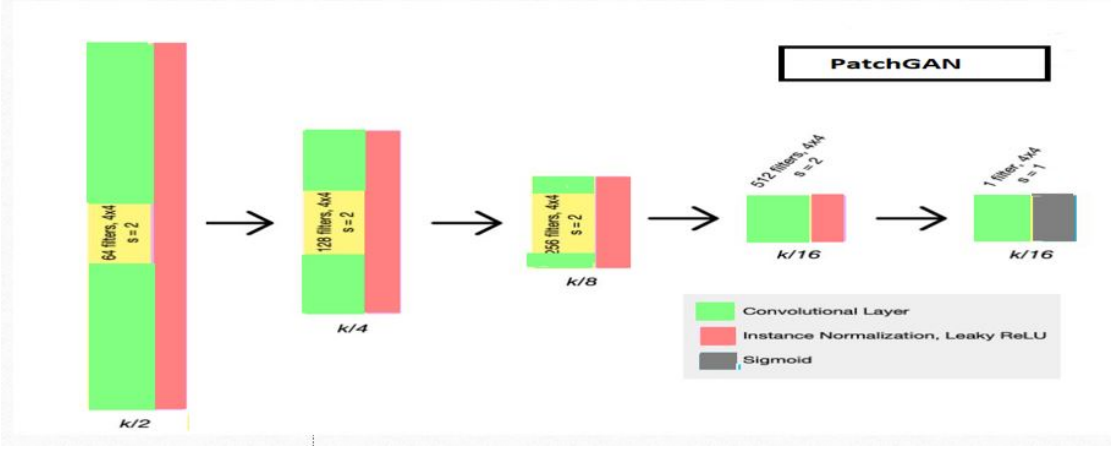

FIGURE 3.3: The patchGAN discriminator architecture

The model additionally comprises a final hidden layer C512 with a 1x1 stride and an output layer C1 with a linear activation function. (In the architecture, Ck represents a 4×4 Convolution-InstanceNorm-LeakyReLU layer with k filters and a stride of 2).

3.2.2 Loss Functions

Our goal is to educate the network translation of visible face image into thermal face image while maintaining the structure of the visible face image in the synthesized image. Therefore, two loss functions are needed.

3.2.3 Adversarial Loss

The adversarial loss [35] also known as GAN's loss is used to urge the output to be visually identical to photos from the target domain. The loss is as follows:

$$L_{GAN}(G, D, X, Y) = E_{y \sim Y} log D(Y) + E_{x \sim X} log(1 - D(G(x)) \quad (3.1)$$

The adversarial loss was firstly used in traditional generative adversarial networks and is a helpful tool for training an unsupervised generative model. The model first requires training a discriminator model to distinguish between actual (dataset) and fake (generated) images and then utilizing the discriminator to train the generator model. The generator will then be updated to encourage it to produce bogus images that are more likely to mislead the discriminator. The discriminator is a binary classifier trained using a binary cross-entropy loss function. But the drawback of this loss function is that it is more concerned with whether or not the predictions are correct rather than how correct or incorrect they may be. For example, when we utilize bogus samples to upgrade the generator model by convincing the discriminator that they are from actual data, there will be nearly no errors because they are on the right side of the decision border, i.e., the right data side. To tackle the drawback of the adversarial loss, the least squares loss is used to substitute the negative log-likelihood objective of the adversarial loss.
Instead of adversarial loss, we utilized the least squares loss. The LSGAN modifies the architecture of GANs, upgrading the discriminator's loss function from binary cross-entropy to least squares loss. Which is as follows:

$$least_{squareGAN} = \sum (y_{predicted} - y_{true})^2 \tag{3.2}$$

This adjustment of the least-squares loss penalizes the generated images based on their distance from the decision boundary. This also addresses the problem of saturation loss by providing a high gradient signal for created images that are significantly distinct or far from the current data. This can be seen in Figure 3.4, which is taken from [51]. The blue line on the left display the sigmoid decision boundary, while on the right the red line displays the least-squares decision boundary. On the right, the pink points represent bogus points that are very far from the decision boundary. so more gradient is applied to move these pink points close to the decision boundary.

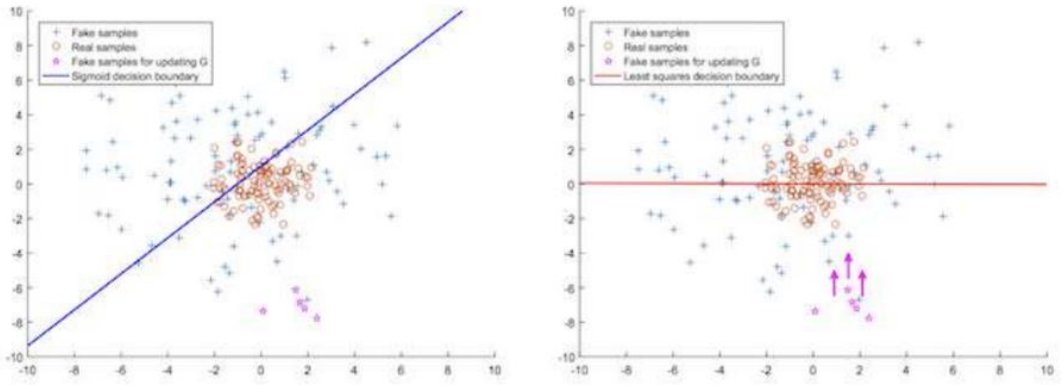

FIGURE 3.4: Comparison between the Least Squares Decision Boundary and the Sigmoid Decision Boundary for Updating the Generator derived from[51].

3.2.4 Patchwise Contrastive Loss

The Patchwise contrastive loss is another loss that is used in our model.The aim of contrastive learning is to maximize the mutual information between input and output by relating the query instance to its positive instance as compared to other instances in the dataset known as negatives.

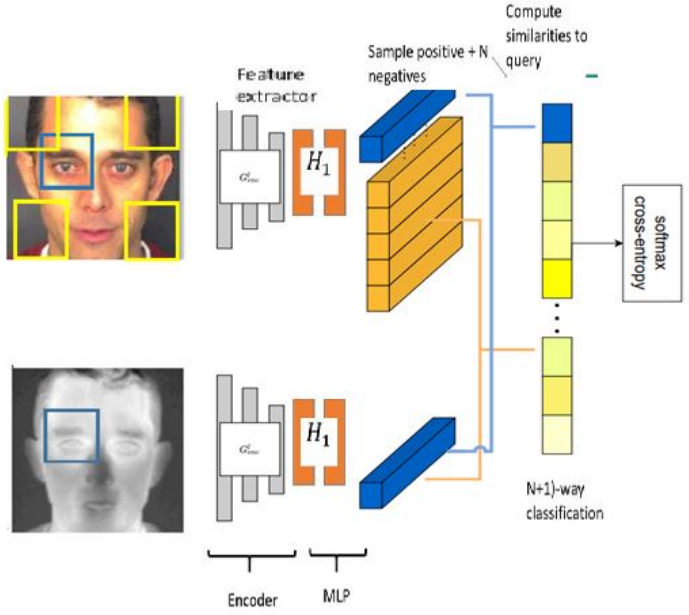

FIGURE 3.5: Workflow of Patch-wise Contrastive loss

Among the source image and target image, a noisy contrastive estimating framework [52] is utilized to maximize mutual information.

To relate the query instance with its positive instance firstly all the instances, including the query, and its positive and negative samples, are mapped into K-dimensional vectors, a a^+ $\epsilon \mathrm{R}^K$, a^- $R^{N\times K}$ and a_n^- R^K denotes the n-th negative.

The normalization is then applied to the embedding space that projects all the z's on the unit sphere and prevents the space from expanding or collapsing.

The problem of (N + 1)-way classification is arranged, where a temperature = 0.07 scales the distances between the query and other instances and is passed as logits [53],[54]. Cross-entropy loss is then determined, describing the likelihood of a positive instance being chosen instead of a negative one as shown in the Figure 3.5.Mathematically contrastive loss is written as:

$$l(a;a^+,a^-) = -log\left[\frac{exp\left(\frac{a;a^+}{T}\right)}{exp\left(\frac{a;a^+}{T}\right) + \left(\sum_{b=1}^{N}\right)exp\left(\frac{a;a^-}{T}\right)}\right]$$

In unsupervised learning, contrastive learning is used at both the image and patch levels [55],[56]. In our task visible to the thermal transformation of facial images, the structure of both the source and target images have typical symmetry, in this case not only the entire image exchange contents but also the matching patches between source and target images. For instance, a patch showing the nose of the output face should be easier to distinguish from the other patches in the face image than the matching nose of the input face. Therefore a patch-based approach is utilized.

In order to perfrom translation , firstly the encoder, G_{enc},, is computed. Its feature set is easily available, and we benefit from it. Within this feature stack, each layer and spatial place depicts the patch of a source image,with deeper layers indicating larger patches.Several Layers of interest are selected, then passed the

feature maps by a 2-layer MLP network H_l, generating a pyramid of features $(\mathrm{z}_l)_L$ = $(\mathrm{H}_l\ G^l_{enc})(x))_L$.Here G^l_{enc} represents the output of the l-th selected layer. Layers are indexed $l\epsilon(1, 2, 3...L)$ and represented by s $\epsilon(1, ...,\mathrm{S}_l)$, where S_l represents the total of spatial locations in every layer. z^S_l. refers to the corresponding feature, and $z^{S|s}_l$. refers to the uncorresponding features. Similarly, the output image is also encoded in the same way as $(\mathrm{z}_l)_L$ = $(\mathrm{H}_l\ G^l_{enc})(x))_L$. The PatchNCE loss is defined as:

$$l_{patchNCE}(G, H, X) = E_{x\sim X} \sum_{l=1}^{L} \sum_{s=1}^{S} l(\hat{Z}_l, z^S_l, z^{S|s}_l) \tag{3.4}$$

The ultimate goal is that the created image must be realistic, and patches in the output and input images must match. The loss of PatchNCE $L_{PatchNCE}(G, H, Y)$ on domain Y images is also utilized to restrict the generator from constituting needless changes. Such loss was effectively specific to a domain, a trainable alternative to identity loss, which has previously been used for unpaired translation techniques [60],[1]. The final loss is as follows:

$$L_{GAN}(G, D, X, Y) + \lambda_X L_{PatchNCE}(G, H, X) + \lambda_Y L_{patchNCE}(G, H, Y) \tag{3.5}$$

3.3 Summary

This chapter covered the methodology in detail. It contained detail about the background, architectures used, and the loss functions. Patchwise loss with the conjunction of the adversarial loss is used to generate a thermal face image from its corresponding visible face image.

Chapter 4

Results and Discussion

The following section discusses the results and discussion in detail.

4.1 Datasets

The most crucial component of machine learning, and artificial intelligence is data. Without data, we are unable to train any model, and all current automation and research are ineffective.
The databases that are utilized in this thesis for the visible to thermal transformation of facial images are as follows:

4.1.1 Carl Database

The Carl database is used in this work which contained face images of three different modalities. This database was developed by Duro et al. [2], and consists of visible, near-infrared, and thermal pictures taken simultaneously under various lighting circumstances as shown in the Figure 4.1.
Each participant supplied five distinct images throughout four collection sessions, using three different picture sensors and lighting setups. This translates into a total of 7.380 pictures (41x4x5x3x3).

Each user was recorded during four distinct acquisition sessions between November 2009 and January 2010.In this context, changes in some people's haircuts and facial hair may be appreciated. The acquisitions were completed throughout the day from 9 AM to 5 PM. A skilled user's whole acquisition process has taken an average of 10 minutes, whereas a non-skilled user's process has taken an average of 15 minutes. Each session took two days to obtain the entire set of users.

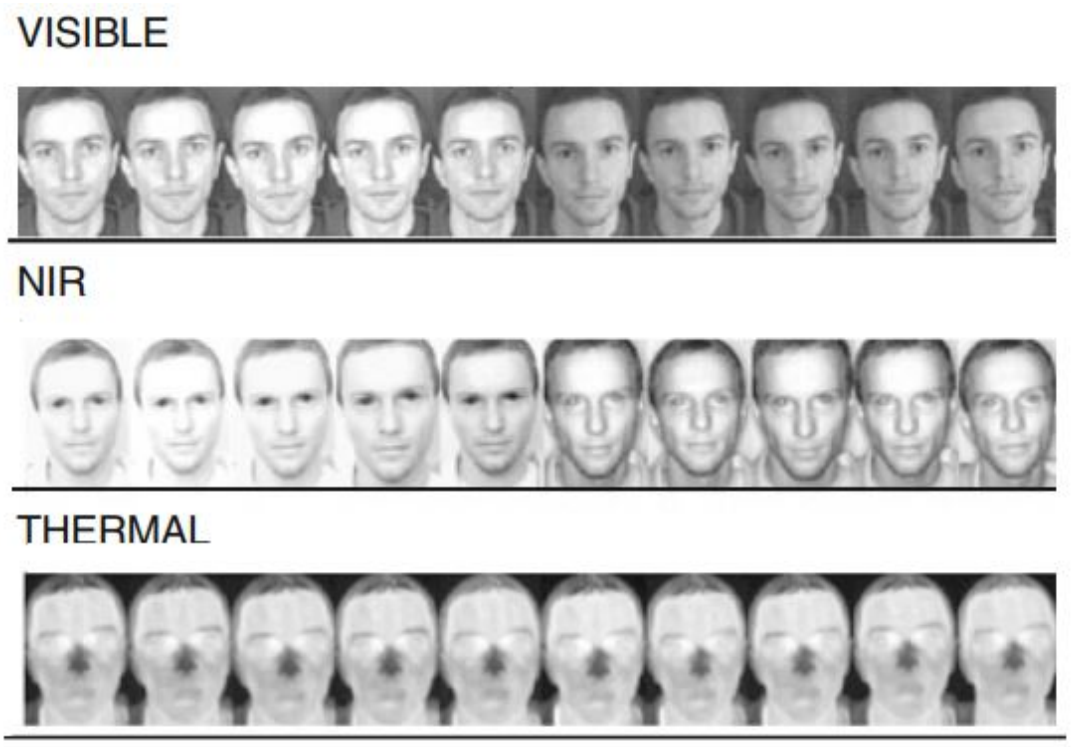

FIGURE 4.1: Example images taken from the Carl database [2]

Five separate frontal images were taken for each illumination condition. Before a facial photograph was taken, anyone who wore glasses was requested to take them off. To obtain an image, no other physical constraints were considered.
Three different illumination setups were used in every session to get the photos (natural, near-infrared, and artificial). Natural illumination (NI) was achieved by opening the windows and allowing sunshine to enter the room. This illumination does not remain constant throughout the day (because of weather conditions). Additionally, it changes based on the time of day. To achieve infrared illumination graphical user interface was developed to allow the intensity level of the IRED

and other picture-related properties to be configured appropriately (brightness, gamma, and exposure).Also to create the artificial illumination of the scene a set of 9 cool fluorescents uniformly dispersed tubes were used. To smooth and fill the infamous disrupted fluorescent spectrum radiation and provide additional IR light, the second pair of Tungsten halogen 650W-3,400K IANIRO Lilliput lights was utilized.

After taking the pictures a Viola and Jones face detector [9] was utilized to eliminate the background and normalize all the pictures to the same size. Although, the Viola and Jones face detector were not able to correctly segment the thermal images. promoting the development of the newly segmented faces [52] method to obtain thermal pictures. All faces were segmented and then scaled using Bi-cubic interpolation to 100 x 145 pixels. The thermal pictures were first saved in the TESTO company's *.bmt format. The metadata in this *.bmt file includes VIS pictures, a temperature matrix, and information of the outside humidity. After processing, the VIS spectrum image was retrieved from this file. The temperature matrix was then saved as a *.mat file in MATLAB, and vthe isible picture was onverted to a grayscale image and saved as a *.bmp file.

Since the Carl face dataset contained images of 3 modalities including near-infrared, visible and thermal and 41 identities. For the experimentation, visible and thermal spectrum images are selected from these. The chosen images are then divided into train and test folders (equated to 1400 paired images of 35 identities for training and 286 paired images of 6 identities for testing).

4.1.2 Tuft Face Database

The Tufts Facial [3] database has almost 10,000 photos with various image modalities, including near-infrared, visible, LYTRO, thermal, recorded video, and 3D images. There are 74 females and 38 males, ages 4 to 70, and more than 15 various nationalities.

To obtain images each subject was seated close to the camera and in front of a blue background. In order to ensure the optimum image center, the cameras were mounted on tripods and their heights were manually adjusted. During the

acquisition process, it was important to carefully control the distance between the camera and the subjects. Further, to provide a consistent lighting environment, diffuse lighting was used.

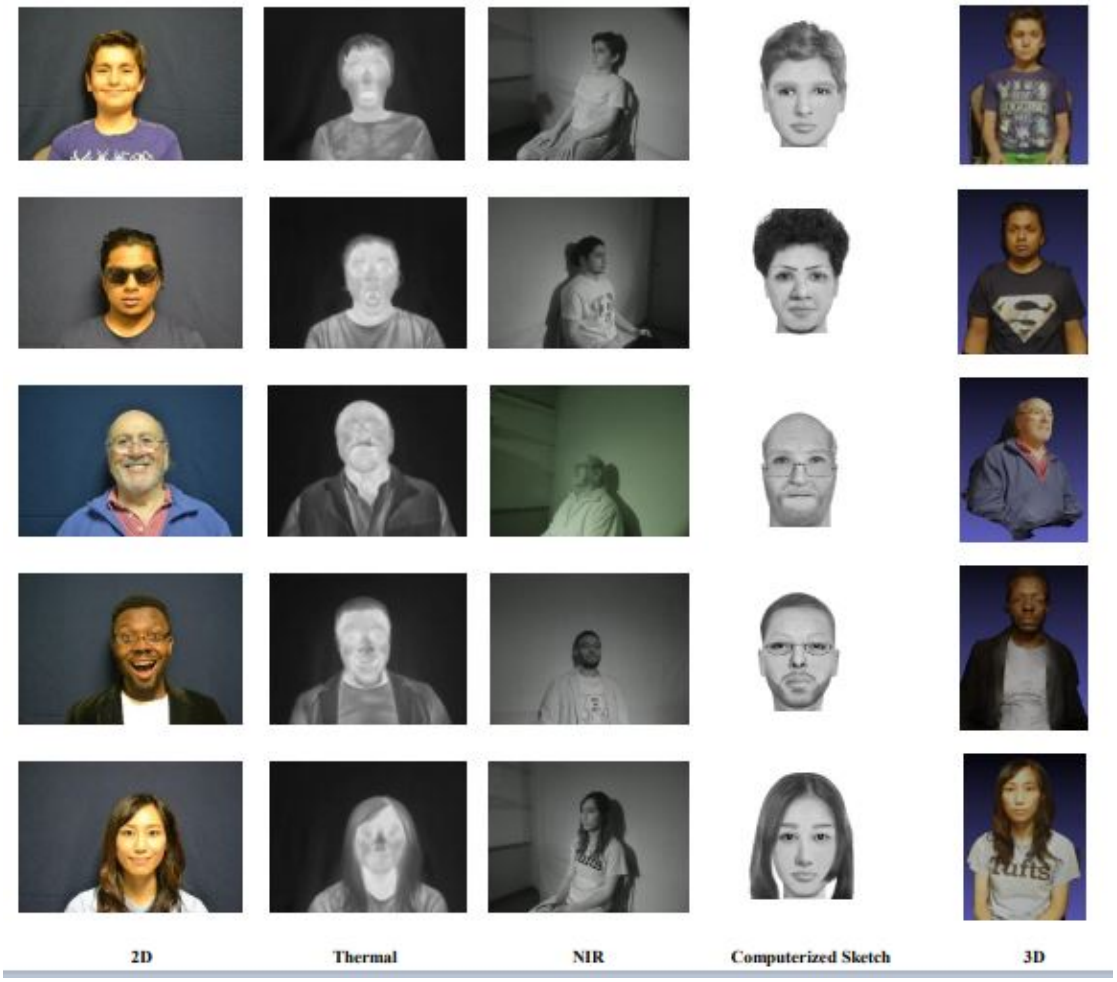

FIGURE 4.2: Images which are taken from the Tufts Face Database an example [3]

Figure 4.2 displays the various image modalities included in this database. In order to obtain 3D images, four cameras were utilized. Each participant was asked to fix their attention on a fixed point while the camera was moved at nine uniformly spaced locations to create an approximate semicircle around them. Then to recreate the 3D models, structure-from-motion methods were utilized.
Computerized facial sketches were created using the FACES 4.0 [57] software. U This software is among the software packages that the FBI, the US Military, and

law enforcement agencies utilize the most. Using this software, researchers can choose candidates from the dataset in accordance with their observations or memories.

Since the tuft face dataset contained images of six modalities including near-infrared, visible, LYTRO, thermal, recorded video, and 3D images and 113 identities. For the experimentation, visible and thermal spectrum photos are selected from these. The chosen pictures are then divided into train and test folders (equated to 800 paired images of 103 individuals for training and 86 paired images of 10 individuals for testing).

4.2 Evaluation Metrics

To evaluate our model four evaluation metrics are utilized .Its details are as follows:

4.2.1 FID

The Frechet Inception Distance [58] (FID) is a statistic to compare the feature vectors generated for real and bogus images. The value represents how similar the two datasets are in computer vision statistics derived using the inception v3 image classification model. Lower values indicate that the two datasets are more comparable or similar in statistics, whereas an FID score of 0.0 implies that the two datasets are identical. To calculate the FID score, The first step is to load a pre-trained Inception v3 model. The model's output layer is then eliminated, and the activations from the last pooling layer are selected. Each image can predict 2,048 activation features because the previous pooling layer includes 2,048 activations. This is referred to as the image's coding vector or feature vector. A 2,048 feature vector is computed for a collection of authentic images from the problem domain to offer a reference for how authentic images are represented. The synthetic images feature vectors are then calculated. In this way groups of 2,048 feature vectors for both actual and generated images will be created. The Frechét inception distance [58] between the activation distributions of the original and synthesized datasets

are computed as:

$$d^2 = \|\mu_1 - \mu_2\|^2 + T_r(C_1 + C_2 - 2\sqrt{C_1 * C_2}) \tag{4.1}$$

In equation (5.1) the terms " μ_1" and "μ_2 " denotes the feature-wise mean of both the authentic and fake images, respectively. The C_1 and C_2 in equation (5.1) represent covari¬ance matrices, often known as sigma, for both synthesized and real feature vectors. The $\|\mu_1$ -$\mu_2\|^2$ represents two mean vectors sum squared difference while T_r stands for the trace linear algebra operation in equation (5.1).

4.2.2 SSIM

The structure similarity index measure [59] (SSIM) is a metric that determines how similar two images are.This metric draws out three key features (Contrast, structure, and Luminance) from the images as shown in Figure 4.3, and based on these features, the two images are compared.

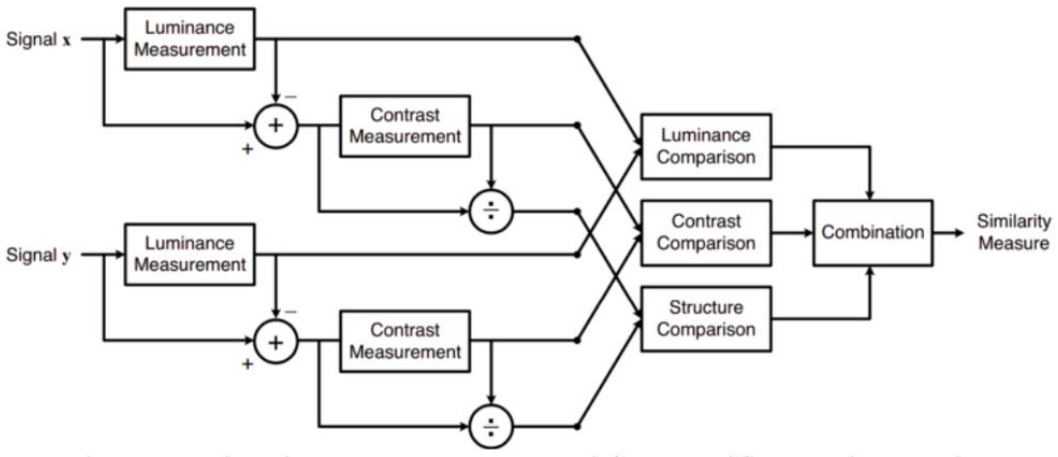

FIGURE 4.3: The (SSIM) Structural Similarity Measurement system's layout and flow from [59]

The SSIM between the target and the generated image is determined by comparing three features (luminance, contrast, and structure) between them. To determine each feature separate function is taken. In the luminance comparison between the

target and the generated image $l(x, y)$ function is used and defined as:

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{4.2}$$

In equation (5.2), μ represent the mean, and is defined as:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{4.3}$$

In equation (5.3), C_1 represents the constant used to stabilize the equation if the denominator becomes 0.

$$C_1 = (K_1 L)^2 \tag{4.4}$$

In equation (5.4) C_1, K_1 are simple constants, and L denotes the range of dynamic pixel values.

In the contrast comparison between the generated and the target image, the function $C(x, y)$ is used and defined as:

$$C(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{4.5}$$

In equation (5.5), σ_x represents the standard deviation and is defined as:

$$\sigma_x = \left(\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_x)\right)^{\frac{1}{2}} \tag{4.6}$$

In the structure comparison between actual and generated images, the function s(x,y) is used and defined as :

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{4.7}$$

Finally, these three features were combined that defined above and calculated the SSIM as:

$$SSIM(x, y) = [l(x, y)]^{\alpha}.[C(x, y)]^{\beta}.[s(x, y)]^{\gamma} \tag{4.8}$$

If , , = 1, C3 = C2/2 is assumed,then the expression is simplified as:

$$SSIM(x,y) = \frac{(2\mu_X\mu_y + C_1)(2\sigma_x\sigma_y + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4.9}$$

4.2.3 PSNR

The Peak sig¬nal to noise ratio [60] (PSNR) displays the intensity variations among the generated image and the target image. If the PSNR score is high, the quality of the produced or artificial image is good. To compute the PSNR [60], firstly, its Mean squared error (MSE) [61] is calculated. MSE basically figures out the average square difference between the authentic and the fake images. MSE identifies the typical discrepancy among pixels across the authentic and fake images. The larger the MSE, the greater the contrast between the natural and synthesized photos. The MSE value is estimated as:

$$MSE = \frac{\sum_{M,N}[l_1(m,n] - l_2(m,n)]^2}{M * N} \tag{4.10}$$

In equation (5.10) M and N represent the number of rows and columns in the input image. After the MSE value, the PSNR [60] value is then calculated. The PSNR value is often measured in decibels (dB). PSNR is a rough approximation of how well humans perceive a reconstruction. The PSNR value is estimated using the MSE as:

$$PSNR = 10log_{10}\left(\frac{R^2}{MSE}\right) \tag{4.11}$$

In equation (5.11), R represents the maximum variance throughout the input data. In the case of an 8-bit unsigned integer, R is 255.

4.2.4 UQI

The universal image quality index (UQI) [62] metric was first used to compute image quality measurements for the human visual system (HVS).UQI is expressed

as:

$$UQI = \frac{1}{M}\sum_{j=1}^{M} Q_j \tag{4.12}$$

In equation (5.12) Q_j denotes the local image quality index, and M stands for the total number of steps. Q has an amplitude between [-1,1]. The ideal value of Q is 1, which can only be obtained if the synthesized image matches to its target image. The Q combines three elements to de¬scribe any distortion: correlation loss, luminosity distortions, and contrasting distortions. The Q is written mathematically as the combination of three expressions:

$$Q = \frac{\sigma_{xy}}{\sigma_x \sigma_y} . \frac{\bar{x}\bar{y}}{(\bar{x})^2 + (\bar{y})^2} . \frac{2\sigma_x \sigma_y}{\sigma_x^2 \sigma_y^2} \tag{4.13}$$

In equation (5.13) , the first expression represents the correlation coefficient between the target and synthesized image. Its value ranges from -1 to 1. The second expression, which has a value range of 0 to 1, determines how near the target and the synthesized image are in terms of mean luminance. The third expression, which also has a range of 0 to 1, measures how close both images are in terms of contrast.

4.3 Preprocessing with Retinaface

In order to train the CUTV2T model, we used two facial databases (Carl and Tuft). Both of them contain visible face images and thermal face images, although they are not aligned and nor have the same size. To achieve good results image alignment is very important. So to make all the visual face images and thermal face images of the same size and aligned, we utilized the Retinaface [63] algorithm. Using RetineFace [63] , you may accomplish two jobs in one shot, including face detection and 2D facial alignment. There is only one consideration made when solving the two distinct targets: that each point in the regressed data for each job should be on an image plane. The model comprises three primary parts: the Context Head Module, the Feature Pyramid Network, and the and the Cascade Multi-Task Loss as shown in the Figure.

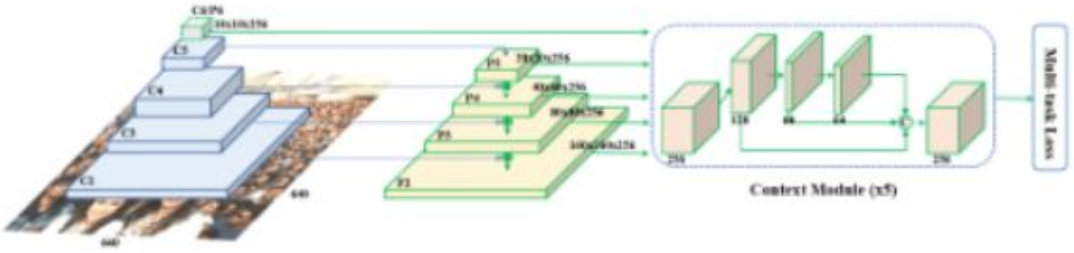

FIGURE 4.4: Architecture of the RetinaFace model [63]

Using the input image as a starting point, the Feature Pyramid Network generates five feature maps of various scales. ResNet's architecture, which was pre-trained on an Imagenet dataset of 11k pictures, was used to construct the first four feature maps in the Figure. The top-most feature map was produced by applying a 3x3 convolution and stride 2 to C5.

To improve the context modeling capacity, a convolutional deformation network is employed in this module instead of a standard 3x3 convolution. Multitask loss and cascading regression are combined to enhance facial localization. Regular anchors forecast the bounding box in the first context module, and regressed anchors are utilized in the following modules for more accurate predictions.

4.4 Training Details

The default settings of CycleGAN [1] except the contrastive loss replaced the l1 cycle-consistency loss is used, including the ResNet Generator containing nine residual blocks, basic PatchGAN discriminator, the least squares GAN loss, Adam optimizer, learning rate of 0.002, and a batch size of 1.

Our model is trained up to 200 epochs. The first half of the CycleGAN generator represents our encoder network, and features are extracted from five layers to compute the patch based contrastive on the multiple layers. Moreover, we sampled 256 random locations for each layer's features and applied two layer MLP network to get 256-dim final features. Following MoCo's setup, we adjusted the momentum to 0.999 and the temperature to 0.07 for our model. Additionally, within each

iteration, we enqueued 256 patches for each image.

TABLE 4.1: The quantitative results with different parameter settings

	Training settings					Evaluation results on carl database			
Method	Id	Negs	Layers	int	Ext	FID↓	SSIM ↑	UQI ↑	PSNR ↑
CUTV2T(even)	✓	255	All	✓	⋊	115.11	0.40	0.72	12.57
CUTV2T(odd)	✓	255	All	✓	⋊	63.49	0.73	0.88	23
Ext only	✓	255	All	⋊	✓	72.70	0.65	0.84	15.25
Last	✓	255	last	✓	⋊	78.18	0.49	0.85	16.13
no id	⋊	255	All	✓	⋊	64.89	0.59	0.86	18.5

Below, some experiments have been established to achieve the results, identify patterns, and gain a deeper understanding of our model.

4.4.1 Experiments with Different Parameter Settings.

To gain a deeper understanding of the proposed CUTV2T model. We did five experiments on the Carl database as shown in the table 4.1 :

- To perform the first experiment training setup outlined in [64] , which includes negative taken from the input image or source image, the PatchNCE loss on domain Y, and the encoder with five layers that derive the features from five evenly distributed points of the encoder (0,4,8,12,16) are used.

- In the second experiment, a few adjustments have been made to the training setup of experiment 1. The multilayer features are extracted from five odd distributed points (1,5,7,9,13) of the encoder and the value of the NCE and GAN is selected as 1.5.

- In the third experiment, the training setup of the second experiment is followed, except that the negatives are sampled from a mini-batch.

- In the fourth experiment, the training setup of experiment 1 is followed, except that the contrastive loss is computed on only the sixteenth layer of the encoder.

- The fifth experiment uses the same training setup as experiment 2, except that identity loss is not taken into consideration.

Following are the test results for each experiment.

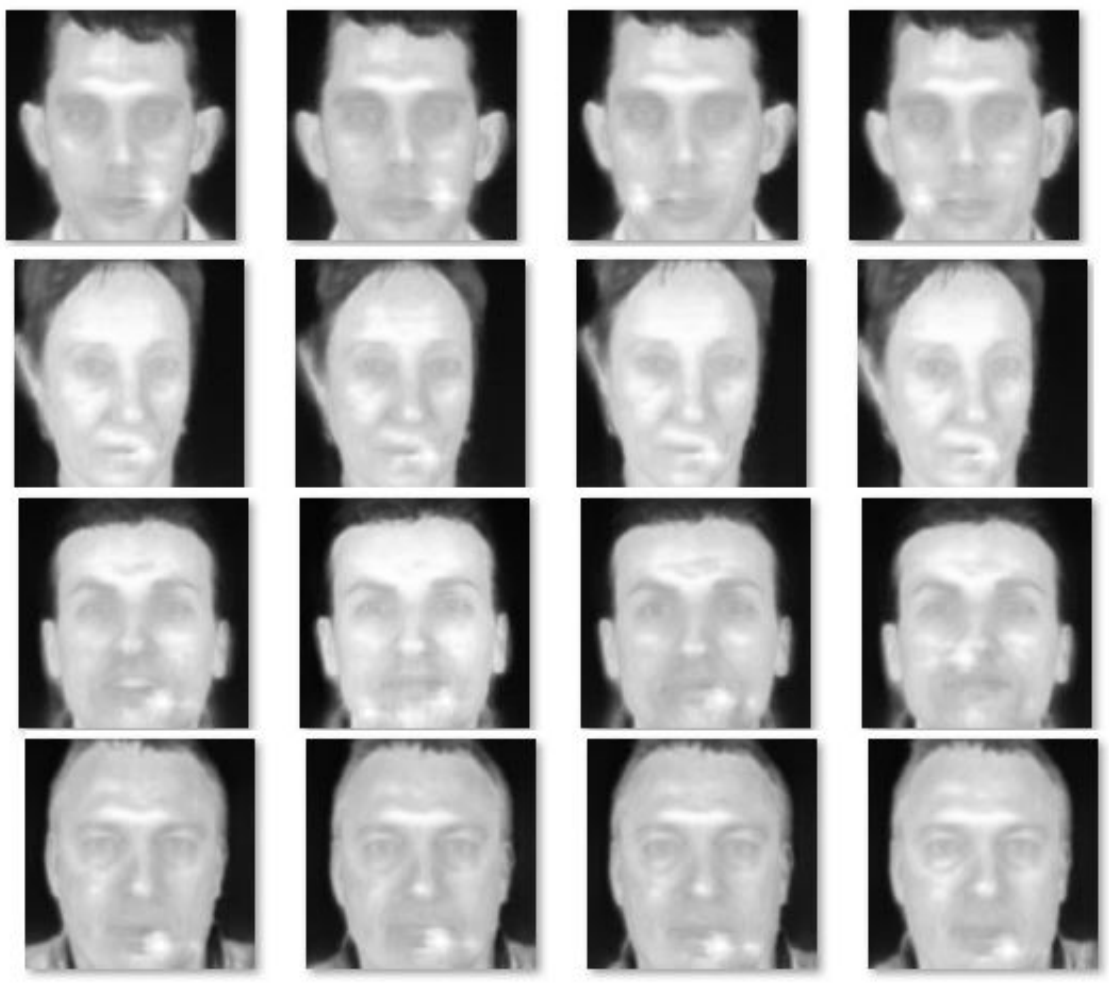

FIGURE 4.5: Translated results for CUTV2T(even).

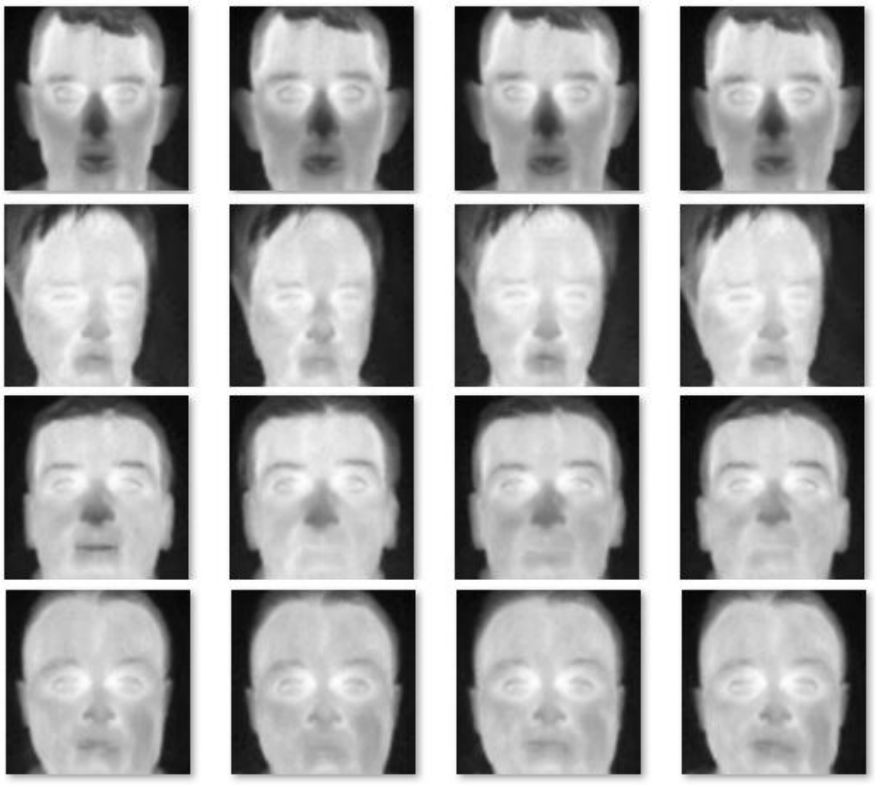

FIGURE 4.6: The translated results for CUTV2T(odd)

FIGURE 4.7: The translated results for Ext only.

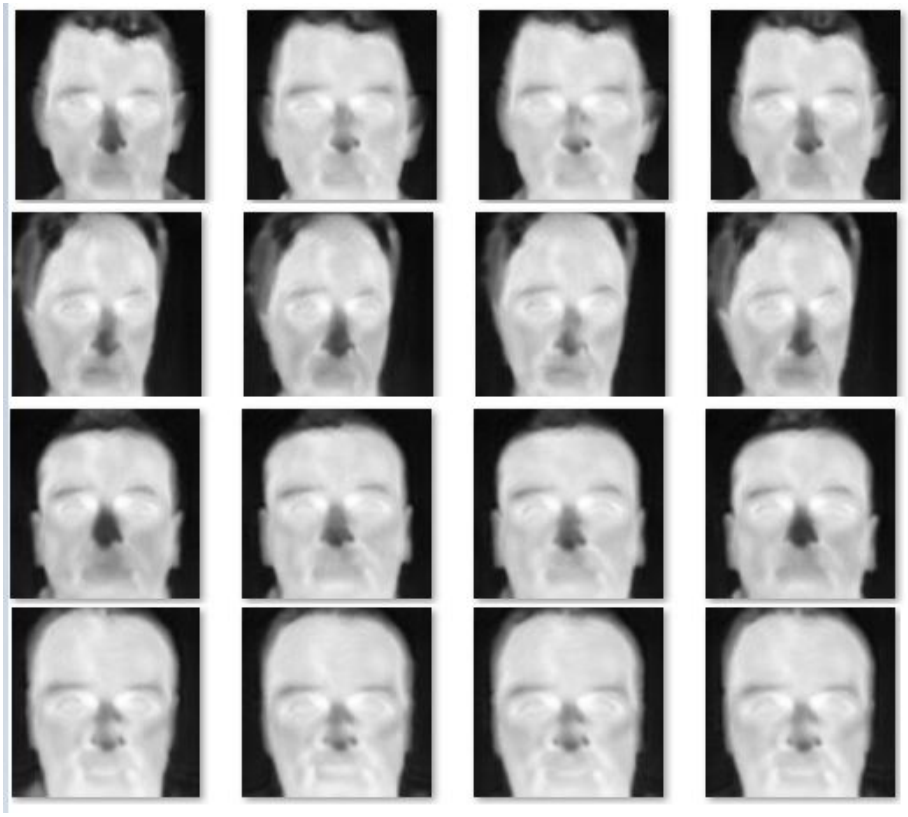

FIGURE 4.8: The translated results for Last.

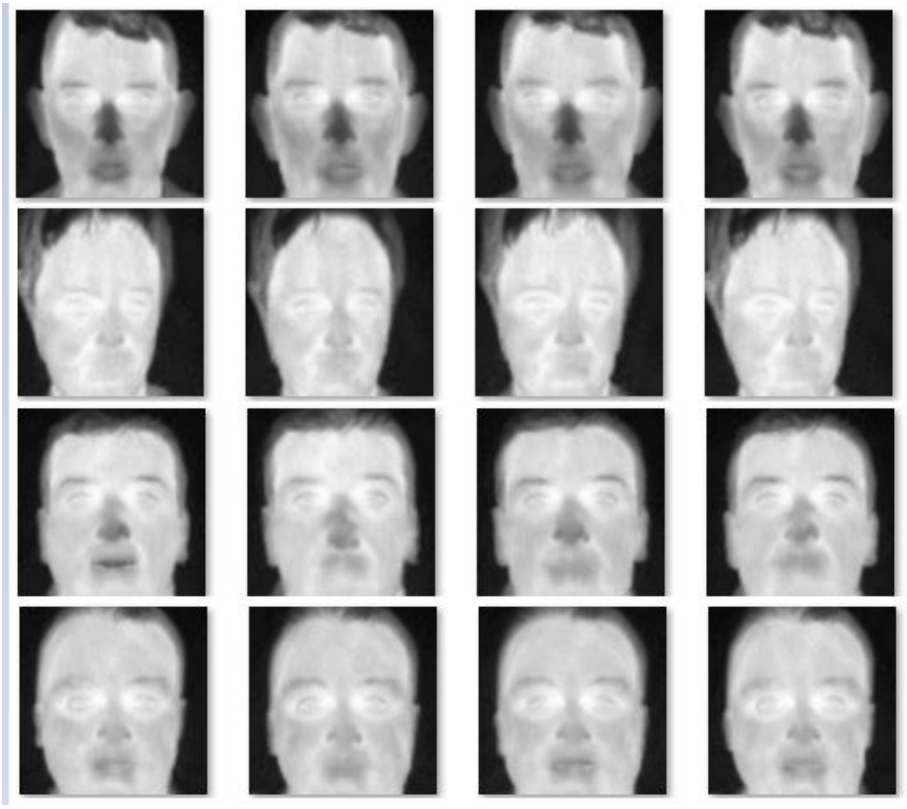

FIGURE 4.9: The translated results for no id.

4.5 Discussion

Table 4.1 indicates the qualitative results with different parameter settings of the proposed CUTV2T on the Carl Face database. The table demonstrated that in the unsupervised image translation settings [65],[53],[66], the implementation configurations of contrastive loss is very important. These configurations include the number of negatives and how to sample them, various choices of hyper-parameter, and data augmentations are all the critical factors to consider and should be well researched.For Example, as shown in the results of the default option of Table 4.1 , in which the contrastive loss is calculated on five evenly distributed points, but the performance is not satisfactory in this case as the parameter setting of contrastive loss is inappropriate for our task visible to thermal transformation of facial images.

During experimentations, we analyzed that finding the correct choices of lambdas for both the NCE and the GAN loss is essential. If a too small values for lambdas are chosen, the model memorizes the training samples or Overfitted. However if a too large value of lambdas are chosen the model is then Underfitted (not effectively learn the training data).

4.5.1 Taking Negatives from within the Same Image is more Powerful

As part of the training process, a variety of training configurations are employed, in which the negatives that are sampled from within the input image are considered as internal negatives and negatives that are derived from other images are considered as external negatives. However, our results have shown that internal negatives produce better results than external negatives when it comes to predicting future outcomes.A stronger signal for content preservation can be gained through taking negatives of the input image, as is shown in the Table 4.1 above.

4.5.2 The Importance of Employing Multiple Encoder Layers

Our model uses an encoder that has multiple layers(1,5,7,9,13). This is compatible with the typical usage of 'l_1+VGG loss,' which employs layers ranging from the pixel level to a deep convolutional layer. Many unsupervised learning studies based on contrastive learning mapped the entire image into a single representation. To imitate this, we used the encoder's last layer (last) and a variant that only uses internal negatives (internal only, final). The performance is not affected too much but training stability is reduced in this case. In our application, the input images are fixed and the loss is used as a signal for image synthesis. Thus, during the image translation process, the dense supervision provided by multiple layers of the encoder is important.

4.5.3 The Regularizer $L_{PatchNCE}(G, H, Y)$ Stabilizes Training

The regularizer with the patch-based contrastive loss urges the generator to leave the image unmodified if the output domain Y is given to the regularizer. The training configurations apart from the regularizer is also tried. The performance does not suffer too much in this case. But with the regularizer, we notice that the training is more stable.

4.5.4 Updating Decoder without PatchNCE

In this training settings, the adversarial loss is only responsible to update the weight of the decoder, in other words, the PatchNCE loss is not used to update the decoder. The results have a lot of artifacts in this case. This demonstrates that the PatchNCE loss not just aids in the learning of the encoder Genc, as has been done previously in unsupervised feature learning approaches [53], but also aids in teaching a superior decoder Gdec in conjunction with the GAN loss. Implicitly, suppose the created result has a lot of artifacts and isn't realistic. In that case, the

encoder will have difficulty finding correspondences between the input and output, resulting in a high PatchNCE loss.

4.5.5 How does our Model Transform the Visible Face Image to the Thermal Face Image?

Our generator model is made up of two networks: an encoder and a decoder. These two networks work together to create a synthetic image in a way that output patches may be easily detected by their input patches. The encoder Genc trains to extract domain-invariant features from the input image, such as hairstyle, eyes, nose, and so on, while our decoder Gdec learns to synthesize domain-specific features of the target domain (thermal images), etc.

4.6 Computational Resources

All tests are conducted on a Tyan server that runs Ubuntu 18.04 LTS and is equipped with Core i7 processors from the tenth generation, 16 GB of RAM, and one Nvidia RTX 3060 Ti GPU.

4.7 Quantitative Comparisons

Based on SSIM, PSNR, UQI, and FID, the quantitative comparisons of the proposed model with the CycleGAN [1] and pix2pix [4] models are made. In table 4.2 and table 4.3, the CUTV2T represents our model trained with identity loss ($\lambda_Y = 1$).The results show that considering FID, UQI, and PSNR metrics proposed CUTV2T model beat both pix2pix and CycleGAN models on Carl and Tuft facial database and achieved the PSNR of 23.05 dB, FID of 63.49, and UQI of 0.88.

TABLE 4.2: Quantitative Comparison with CycleGAN [1] and pix2pix [4] on Carl database [2].

Method	Dataset Used	FID ↓	SSIM ↑	PSNR(dB) ↑	UQI ↑
CUTV2T	Carl dataset	**63.49**	0.73	**23.05**	**0.88**
CycleGAN [1]	Carl dataset	94.75	0.53	16.35	0.85
pix2pix [4]	Carl dataset	130.63	0.77	19.25	0.81

TABLE 4.3: Quantitative Comparison with CycleGAN [1] and pix2pix [4] on Tuft database [3]

Method	Dataset Used	FID ↓	SSIM ↑	PSNR(dB) ↑	UQI ↑
CUTV2T	Tuft database	**85.13**	0.589	**18.03**	**0.87**
CycleGAN [1]	Tuft database	100.75	0.58	14.3	0.85
pix2pix [4]	Tuft database	115	0.72	15.47	0.80

4.8 Qualitative Comparisons

Using two databases (Carl and Tuft face database) the qualitative comparison of the proposed CUTV2T with the CycleGAN [1] and pix2pix [4] has been made.

FIGURE 4.10: Input images from the Carl database[2]

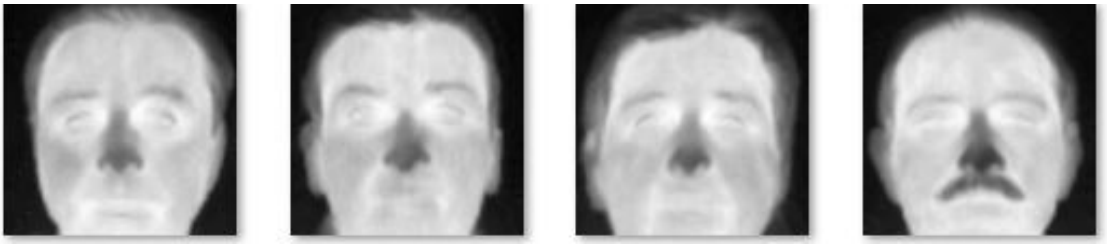

FIGURE 4.11: Results of CUTV2T model on Carl database[2].

FIGURE 4.12: Results of CycleGAN [1] model on Carl database[1].

The qualitative results show that the proposed CUTV2T captured the heat signatures of human faces prominently on both Carl [2] and Tuft [3] databases . The qualitative results on the Carl [2] database and Tuft [3] database are as follows.

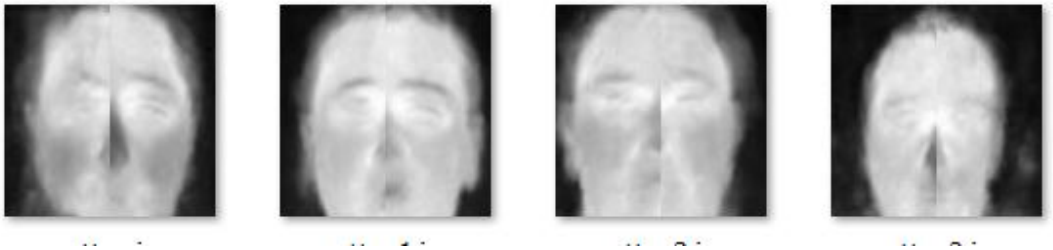

FIGURE 4.13: Results of pix2pix [4] model on Carl database.[2]

Figures 4.11, 4.12 and 4.14 show the results of CUTV2T, CycleGAN [1], and pix2pix [4], models, and all are developed using the Carl dataset [2]. When qualitatively examined, we found that the nose area had the most variation when training CycleGAN [1]. Additionally, In comparison to the CUTV2T model, Pix2pix [4] trained on Carl [2] exhibits structural distortions and smearing with fewer articulations.Similarly, the qualitative comparison of the proposed CUTV2T with the Pix2pix [4] and CycleGAN [1] model is also made as shown in the Figures 4.15, 4.16,4.17.The results indicate that the proposed CUTV2T performs better compared to the CycleGAN and pix2pix.

FIGURE 4.14: Input/source images from Tuft database [3].

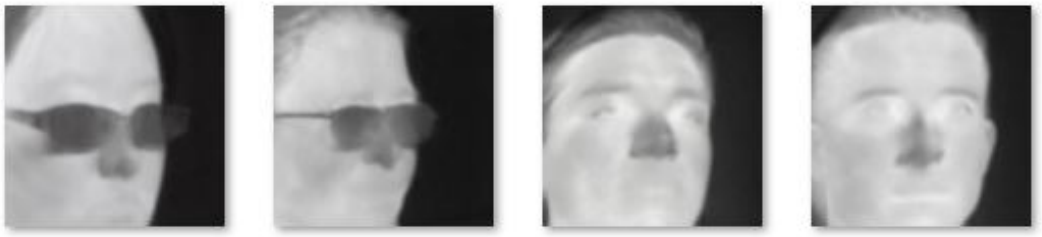

FIGURE 4.15: The proposed CUTV2T results on the Tuft facial database [3].

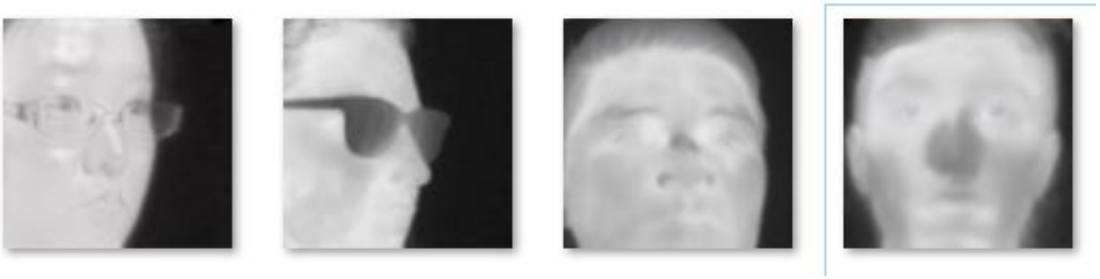

FIGURE 4.16: Results from the CycleGAN [1] using the Tuft face database[3] .

FIGURE 4.17: Results from the Pix2pix [4] using the Tuft face database [3].

4.9 Summary

This chapter provided the detail of the results and discussion. Two thermal visible face databases (Carl database [50] and Tuft database [51]) were utilized in order to perform experimentations. The quantitative and qualitative comparisons were also included in this chapter.

Chapter 5

Conclusion and Challenges

5.1 Conclusion

In this study a new model for visible to thermal transformation of facial images has been proposed, that can train on unpaired facial images. It has been shown that the proposed CUTV2T is an excellent candidate for this transformation. The proposed CUTV2T model has shown very good results as evaluated considering the PSNR, UQI, and FID score and achieved the PSNR of 23.05 dB, FID of 63.49, and UQI of 0.88. The comparison has also been made with the other well-known translation model including Pix2pix [4] and CycleGAN [1] and shown that the CUTV2T has shown superior performance. The proposed CUTV2T based on CUT [64],uses an objective to develop an embedding that brings corresponding patches in input and output together while pushing non-corresponding "negative" patches away. so it is more suitable for unpaired translation.

5.2 Future Work

The proposed system works on only facial images which have a typical symmetry but it is not designed for general purpose indoor/outdoor images. Further investigation is needed for the transformation of indoor/outdoor images especially using

unpaired data.

5.3 Challenges

The CUT [64], as well as GAN models, are using adversarial algorithms and all of them face a number of significant challenges. All GAN-based models experience a convergence issue due to their adversarial nature. The adversarial loss has competing terms that must be balanced. As far as the original GAN [35] is concerned, the adversarial loss is the loss of the generator against the loss of the discriminator. In the CUT [64], the contrastive loss is also included and must be balanced against the generator loss, in addition to the generator and discriminator balance. In the CUTV2T, additionally, a regularization on target domain Y, is also added. The idea behind the regularizer-based CUTV2T model is to learn just the required information. Therefore a careful balance between the contrastive loss, the generator loss and the regularization term need to be maintained. In neural networks, such adjustment between loss terms poses complex optimization and convergence issues. In this settings, hyperparameter fine-tuning becomes critical and complex.

Bibliography

[1] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[2] V. Espinosa-Duró, M. Faundez-Zanuy, and J. Mekyska, "A new face database simultaneously acquired in visible, near-infrared and thermal spectrums," *Cognitive Computation*, vol. 5, no. 1, pp. 119–135, 2013.

[3] K. Panetta, Q. Wan, S. Agaian, S. Rajeev, S. Kamath, R. Rajendran, S. P. Rao, A. Kaszowska, H. A. Taylor, A. Samani *et al.*, "A comprehensive database for benchmarking imaging systems," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 3, pp. 509–520, 2018.

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

[5] D. A. Socolinsky and A. Selinger, "A comparative analysis of face recognition performance with visible and thermal infrared imagery," in *Object recognition supported by user interaction for service robots*, vol. 4. IEEE, 2002, pp. 217–222.

[6] J. Ruiz-del Solar, R. Verschae, and M. Correa, "Recognition of faces in unconstrained environments: A comparative study," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 1–19, 2009.

[7] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017, pp. 1–10.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[9] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[10] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[11] K. Regmi and A. Borji, "Cross-view image synthesis using conditional gans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3501–3510.

[12] P. Zhu, R. Abdal, Y. Qin, and P. Wonka, "Sean: Image synthesis with semantic region-adaptive normalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5104–5113.

[13] Q. Yang, N. Li, Z. Zhao, X. Fan, E. I. Chang, Y. Xu *et al.*, "Mri cross-modality neuroimage-to-neuroimage translation," *arXiv preprint arXiv:1801.06940*, 2018.

[14] X. Guo, Z. Wang, Q. Yang, W. Lv, X. Liu, Q. Wu, and J. Huang, "Gan-based virtual-to-real image translation for urban scene semantic segmentation," *Neurocomputing*, vol. 394, pp. 127–135, 2020.

[15] R. Li, W. Cao, Q. Jiao, S. Wu, and H.-S. Wong, "Simplified unsupervised image translation for semantic segmentation adaptation," *Pattern Recognition*, vol. 105, p. 107343, 2020.

[16] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 1857–1865.

[17] Z. Yi, H. Zhang, P. Tan, and M. Gong, "Dualgan: Unsupervised dual learning for image-to-image translation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2849–2857.

[18] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2536–2544.

[19] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," *Advances in neural information processing systems*, vol. 30, pp. 2–5, 2017.

[20] Y. Song, C. Yang, Z. Lin, X. Liu, Q. Huang, H. Li, and C.-C. J. Kuo, "Contextual-based image inpainting: Infer, match, and translate," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[21] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, "A tutorial on energy-based learning," *Predicting structured data*, vol. 1, pp. 1–14, 2006.

[22] J. Xu, H. Li, and S. Zhou, "An overview of deep generative models," *IETE Technical Review*, vol. 32, no. 2, pp. 131–139, 2015.

[23] A. Oussidi and A. Elhassouny, "Deep generative models: Survey," in *2018 International Conference on Intelligent Systems and Computer Vision (ISCV)*. IEEE, 2018, pp. 1–8.

[24] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5485–5493.

[25] H.-M. Chu, C.-K. Yeh, and Y.-C. F. Wang, "Deep generative models for weakly-supervised multi-label classification," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 400–415.

[26] M. Tschannen, E. Agustsson, and M. Lucic, "Deep generative models for distribution-preserving lossy compression," *Advances in Neural Information Processing Systems*, vol. 31, pp. 2–10, 2018.

[27] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[28] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International conference on machine learning*. PMLR, 2014, pp. 1278–1286.

[29] T. Zhang, A. Wiliem, S. Yang, and B. Lovell, "Tv-gan: Generative adversarial network based thermal to visible face recognition," in *2018 international conference on biometrics (ICB)*. IEEE, 2018, pp. 174–181.

[30] Z. Wang, Z. Chen, and F. Wu, "Thermal to visible facial image translation using generative adversarial networks," *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1161–1165, 2018.

[31] H. Zhang, V. M. Patel, B. S. Riggan, and S. Hu, "Generative adversarial network-based synthesis of visible faces from polarimetrie thermal faces," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 100–107.

[32] H. Zhang, B. S. Riggan, S. Hu, N. J. Short, and V. M. Patel, "Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks," *International Journal of Computer Vision*, vol. 127, no. 6, pp. 845–862, 2019.

[33] L. Song, M. Zhang, X. Wu, and R. He, "Adversarial discriminative heterogeneous face recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[34] A.-C. Guei and M. A. Akhloufi, "Deep generative adversarial networks for infrared image enhancement," in *Thermosense: Thermal Infrared Applications XL*, vol. 10661. International Society for Optics and Photonics, 2018, p. 106610B.

[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, pp. 1–10, 2014.

[36] D. Berthelot, T. Schumm, and L. Metz, "Began: Boundary equilibrium generative adversarial networks," *arXiv preprint arXiv:1703.10717*, 2017.

[37] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[38] C. Ordun, E. Raff, and S. Purushotham, "Generating thermal human faces for physiological assessment using thermal sensor auxiliary labels," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 1319–1323.

[39] V. V. Kniaz, V. A. Knyaz, J. Hladuvka, W. G. Kropatsch, and V. Mizginov, "Thermalgan: Multimodal color-to-thermal image translation for person re-identification in multispectral dataset," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 5–10.

[40] P. Liu, F. Li, S. Yuan, and W. Li, "Unsupervised image-generation enhanced adaptation for object detection in thermal images," *Mobile Information Systems*, vol. 2021, pp. 2–8, 2020.

[41] K. K. Babu and S. R. Dubey, "Pcsgan: perceptual cyclic-synthesized generative adversarial networks for thermal and nir to visible image transformation," *Neurocomputing*, vol. 413, pp. 41–50, 2020.

[42] G. Hermosilla, D.-I. H. Tapia, H. Allende-Cid, G. F. Castro, and E. Vera, "Thermal face generation using stylegan," *IEEE Access*, vol. 9, pp. 80 511–80 523, 2021.

[43] V. Pavez, G. Hermosilla, F. Pizarro, S. Fingerhuth, and D. Yunge, "Thermal image generation for robust face recognition," *Applied Sciences*, vol. 12, no. 1, p. 497, 2022.

[44] V. Kniaz, V. Gorbatsevich, and V. Mizginov, "Thermalnet: a deep convolutional network for synthetic thermal image generation," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, p. 41, 2017.

[45] J. Li, P. Hao, C. Zhang, and M. Dou, "Hallucinating faces from thermal infrared images," in *2008 15th IEEE International Conference on Image Processing*. IEEE, 2008, pp. 465–468.

[46] M. Dou, C. Zhang, P. Hao, and J. Li, "Converting thermal infrared face images into normal gray-level images," in *Asian Conference on Computer Vision*. Springer, 2007, pp. 722–732.

[47] K. Mallat and J.-L. Dugelay, "Facial landmark detection on thermal data via fully annotated visible-to-thermal data synthesis," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2020, pp. 1–10.

[48] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[49] S. Gross and M. Wilber, "Training and investigating residual nets, facebook ai research, ca," *Avilable: http://torch. ch/blog/2016/02/04/resnets. html*, 2016.

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

[51] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[52] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[53] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[54] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3733–3742.

[55] P. Bachman, R. D. Hjelm, and W. Buchwalter, "Learning representations by maximizing mutual information across views," *Advances in neural information processing systems*, vol. 32, pp. 2–10, 2019.

[56] O. Henaff, "Data-efficient image recognition with contrastive predictive coding," in *International Conference on Machine Learning*. PMLR, 2020, pp. 4182–4192.

[57] H. K. Galoogahi and T. Sim, "Face sketch recognition by local radon binary pattern: Lrbp," in *2012 19th IEEE International Conference on Image Processing*. IEEE, 2012, pp. 1837–1840.

[58] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, pp. 1–19, 2017.

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

[60] D. Poobathy and R. M. Chezian, "Edge detection operators: Peak signal to noise ratio based comparison," *IJ Image, Graphics and Signal Processing*, vol. 10, pp. 55–61, 2014.

[61] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE signal processing magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[62] ——, "A universal image quality index," *IEEE signal processing letters*, vol. 9, no. 3, pp. 81–84, 2002.

[63] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-shot multi-level face localisation in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5203–5212.

[64] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *European Conference on Computer Vision*. Springer, 2020, pp. 319–345.

[65] J. H. Olivier, R. A. Hénaff, C. Doersch *et al.*, "Data-efficient image recognition with contrastive predictive coding," *arXiv preprint arXiv:1905.09272*, 2019.

[66] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.