Towards Improving the Performance of Multi-Spectral Face Recognition Systems through Image Synthesis

by

Suha Reddy Mokalla

(Under the Direction of Thirimachos Bourlai)

Abstract

Face biometrics has been gaining traction among the biometrics research community in the recent decades due to its many advantages an image can be captured in a covert manner, at a distance, and in-the-wild scenarios; no contact is needed with the sensor, to name a few. Since the inception of convolutional neural networks, many face recognition (FR) models and its related components have been proposed. This is possible because of the availability of many large-scale visible band face datasets. However, models based on visible band data are not reliable in poorly illuminated areas. To address this problem, this dissertation proposes a synthesis-based approach where visible band face images are synthesized from their thermal counterparts to facilitate the usage of the existing state of the art visible band-based FR models. The contributions of this work are three-fold. First, to the best of our knowledge, the world's largest yet dataset of its kind, MILAB-VTF(B) dataset has been collected and annotated at UGA by our team to advance the field of cross-spectral FR.

One of the essential pre-processing steps for many FR systems is geometric normalization, which involves rotating the face in an image so that the line joining the geometrical eye centers is horizontal. For the majority of the face matching algorithms, detecting the eye centers is an important module of the face normalization process. The proposed method improved the eye center detection accuracy by 60% over the baseline model, and by 14% over training only the StarGAN2 model without the alignment loss.

To address the problem of cross-spectral FR, involution, an operation that inverts the inherence of convolution is used as the atomic operation to implement a GAN model that includes a generator, a discriminator, and a style encoder module. Additionally, an identity loss is introduced to preserve the distinguishable characteristics of individual subjects which improved the face verification area under curve (AUC) by around 4% over the current benchmark and reduced the equal error rate (EER) by 7%. On the more challenging MILAB-VTF(B), the AUC is increased by around 16% and 10% over the indoor and outdoor datasets, respectively.

Index words: [Cross-spectral face recognition, Long wave infrared, Mid-wave infrared, thermal spectrum, visible spectrum, eye center detection, geometric normalization,

Involution GAN, StarGAN2, CycleGAN, generative adversarial networks, image synthesis, MWIR, LWIR.]

Towards Improving the Performance of Multi-Spectral Face Recognition Systems through Image Synthesis

by

Suha Reddy Mokalla

B.S., Jawaharlal Nehru Technological University, India, May 2015
M.S., West Virginia University, 2020

A Dissertation Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the Degree.

Doctor of Philosophy

Athens, Georgia

2023

Towards Improving the Performance of Multi-Spectral Face Recognition Systems through Image Synthesis

by

Suha Reddy Mokalla

Major Professor: Thirimachos Bourlai

Committee: Fred Beyette
Mark Haidekker
Wenzhan Song

Electronic Version Approved:

Ron Walcott
Dean of the Graduate School
The University of Georgia
August 2023

Dedication

This dissertation is dedicated to my late grandfather, Mr. Raja Reddy Panyala, my inspiration and the source of strength, and to my late uncle, Mr. Dhananjaya Reddy Panyala.

Acknowledgments

I would like to express my deepest appreciation to my advisor and mentor, Dr. Thirimachos Bourlai, for his support and guidance. I am deeply indebted to UGA and MILAB for providing me with this amazing opportunity and supporting and funding my doctoral program. I am extremely grateful to UGA and the school of ECE for making my transition from my former school effortless. I would also like to extend my deepest gratitude to my family and friends, who supported me in rain and shine.

I would like to extend my deepest thanks to my advisory committee, Dr. Mark Haidekker, Dr. Wenzhan Song, Dr. Fred Beyette and Dr. Mable Fok for their valuable guidance and feedback.

Finally, this work was partially supported by an STTR Phase II contract W911QX20C0022 from the US Army Research Laboratory, Adelphi, MD.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In the last decade, face recognition has been gaining a lot of traction among the biometric research community. Biometric system function by comparing the unique physiological or behavioral characteristics of individuals to perform recognition. Examples of physiological traits are face, fingerprint, palm, iris etc. and examples of behavioral traits are keystroke, gait, voice etc. These systems have a wide range of applications in military, government and non-government fields for access control, banking security etc. Physiological traits have a set of advantages over the behavioral traits as these are permanent, more unique, and vary from one individual to another. Among the physiological traits, imaging systems are more robust than fingerprint and palm since no contact is required with the sensor. Iris recognition has the advantages of it being impermeable genetically, and iris stroma are very unique and are different even in identical set of twins, triplets, etc. Face is considered one of the most important biometric traits in human identification due to the following reasons: data can be collected in a covert manner, at a distance, and in-the-wild-scenarios; no contact is needed with the sensors, to name a few.

On a broader level face recognition can be considered as either a verification or an identification problem. In general terms verification is *"is this the same person?"* or a 1:1 problem, where an image is verified against one registered image. Example of face verification systems are personal devices such as cell phones, computers, storage lockers etc. Face identification can be considered a *"who is this person?"* or a 1:N matching problem. To identify a person means to check if the person exists in a pre-registered database called gallery. The person in the gallery set with the least distance score with the person in question is considered the match. In face recognition, the image that needs to be identified or verified is called a query image, and the registered image or set of images is called gallery.

Face identification results are typically presented as Rank-N accuracy metrics. Rank-1 accuracy is the percentage of predictions where the top prediction matches the ground-truth label. Rank-2 is the cumulative of the top-most and the next top predictions, and rank-3 is the cumulative of top 3. These results can be represented graphically as cumulative match characteristic (CMC) curves, which is rank accuracy plotted against the rank number. Face verification performance is typically presented using area

under curve (AUC) and equal error rate (EER). To obtain these, true positive rate (TPR) and false positive rate (FPR) are calculated over a set of thresholds, and these are plotted against the threshold. The threshold at the intersection of these two lines is called the EER. A region operating characteristic (ROC) curve is typically used to represent the verification results graphically, and this is the plot of the TPR against the FPR. The area under this curve is referred to as the AUC.

Since the inception of neural networks, especially, convolutional neural networks (CNNs) (LeCun, Bengio, et al., 1995), many state-of-the-art (SOTA) face recognition (FR) and related models have been proposed (Deng et al., 2019; Parkhi et al., 2015; Schroff et al., 2015; H. Wang et al., 2018; J. Wang et al., 2020; Xiang & Zhu, 2017). This is due to the availability of several mined datasets in the visible spectrum such as Labeled faces in the wild (LFW) (G. B. Huang et al., 2008), MS-celeb-1M (Guo et al., 2016), VGG-Face2 (Cao et al., 2018), and YouTube Faces (Wolf et al., 2011). However, models based on the visible band data (380-700nm) are not reliable when the data needs to be collected in poorly illuminated areas.

Working outside the visible spectrum, and thus, collecting infrared (IR) face images is proven to be a practical solution to the problem of low light face capturing and matching (Anghelone et al., 2022). IR spectrum is broadly classified into active and passive IR. Active IR is further classified into Short-Wave Infrared (SWIR) and Near Infrared (NIR), while passive IR comprises of Mid-Wave Infrared (MWIR) and Long-Wave Infrared (LWIR). The advantages and disadvantages of these bands have been discussed in (Anghelone et al., 2022). This paper primarily focuses on the data collected in the MWIR (3-7μm) and the LWIR spectra (8-14μm), referred to as the thermal band. It has many advantages over the visible spectrum, for instance, it is invariant to illumination conditions i.e., face image characteristics are found to be unaffected in low-light and no-light conditions. Therefore, no external illumination is required to collect data in the thermal band.

Also, images in most, if not all, of the registered databases are collected in constrained visible conditions, driving the need for cross-spectral FR. The main differences between MWIR and LWIR bands, other than the spectral range are: (1) the LWIR band has a better atmospheric transmission, allowing radiation to pass through the atmosphere with minimal interference; MWIR experiences more attenuation than LWIR while maintaining reasonably good transmission characteristics. (2) LWIR is well-suited for applications that require temperature sensitivity. MWIR offers temperature sensitivity too, but not as pronounced as LWIR. (3) LWIR often requires cooled IR detector and can be more expensive than MWIR. Several bands of the electromagnetic spectrum are shown in Figure 1.2

Most of the FR systems typically comprise of the following building blocks: 1. Data acquisition, 2. Face detection, 3. Geometric Normalization, 4. Feature extraction, and 5. Matching as shown in Figure 1.1. Data acquisition refers to either collecting the data from human subjects or mining data from the internet and other sources. Face detection is localizing the face that needs to be recognized in the given image. Geometric normalization refers to rotating a face image such that the line joining the geometrical eye centers is horizontal. This is also called facial alignment. The next step is extracting the features from the detected and normalized faces, followed by performing the matching.

1.2 Acquisition of thermal-to-visible datasets

This section provides a brief introduction to MILAB-VTF(B) dataset, the largest thermal-to-visible dataset till date, to the best of our knowledge, which is one of the main contributions of this dissertation. This data collection used the latest MWIR imaging sensors, with and without telephoto capabilities, i.e., the A8581 (for close-range) and the RS8500 series (for long-range imaging). The MWIR-visible dataset created from this data collection includes a curated dataset that will be publicly available. Thus, assisting the research community by further closing the gap of MWIR-visible datasets availability. The contributions of this work are as following:

- An unconstrained, multi-spectral (visible and MWIR), unsynchronized, paired face dataset with 400 identities is collected.
- This dataset offers variety in terms of weather, pose, and distances.
- The largest multi-spectral face dataset to date by the number of subjects, distances, and the thermal sensor resolution.

The raw dataset was collected within the 1st quarter of 2021, and is the largest and the most diverse of its kind to-date, collected in realistic operational conditions, from 1.5 - 400 meters (1312 ft.). The data collection activity, curated dataset, and the baseline experiments are explained in section 4.1. Example figures from the dataset are shown in Figure 1.3

1.3 Utilizing Alignment Loss to Advance Eye Center Detection and Face Recognition in the LWIR Band

Geometric normalization is an integral part of most of the face recognition systems and it refers to rotating an image such that the line joining the eye centers is horizontal. This requires determining the accurate geometric eye center location in an image. Reported work in the literature shows the positive effects of eye center localization on the face recognition performance (P. Wang et al., 2005).

While this work focuses on geometric eye center detection, pupil center detection is another research area, which has applications in driver drowsiness detection, human-computer interaction etc. (Fuhl, Santini, et al., 2016; Fuhl, Santini, et al., 2016). This difference is illustrated in Figure 1.4. There are few efficient algorithms developed in the visible spectrum that detect eye centers. One such work is High-Resolution (HR-Net) (Xiang & Zhu, 2017) which detects the face bounding boxes and five landmarks, which are, the eye centers, tip of nose, and mouth corners. These models do not perform well on the images captured outside the visible spectrum, given the spectral gap between the images collected in different bands. One way to geometrically normalize LWIR images is to manually locate the eye centers on each face image that needs to be identified or verified. However, this task becomes impractical when dealing with datasets with thousands or millions of identities as with the real world applications. Geometric

normalization requires an accurate and robust eye center detection algorithm to be employed. Figure 1.4 shows the geometrically normalized images with ground-truth eye center annotations, eye centers detected through baseline method, original StarGAN2 (Choi et al., 2020), and the proposed method.

An expected first solution to this problem would be to develop a new LWIR based eye centers or multi-facial landmark (that includes eye centers) detection model, which would require numerous large-scale LWIR based face datasets that could be used for training. Thus, with a limited dataset, such model would perform poorly (this is demonstrated in the baseline experiments). An alternate solution is to use the existing robust visible band-based landmark detection models to detect the eye centers in the thermal images. The challenge here is the significant differences between the visible and the thermal band face images and thus as tested this approach is reported to not perform well.

To address the aforementioned problem, we propose a thermal-to-visible synthesis based approach to detect the eye centers in thermal images, and utilize an alignment loss derived from the normalized error between the actual and detected eye centers. Two variations of generative adversarial networks (GANs) called StarGAN2 (Choi et al., 2020), and cycleGAN (Zhu et al., 2017) are trained to generate visible images from their thermal counterparts. GANs were first proposed by Goodfellow et al., 2014 and since then many variations of the network have been proposed for various applications. One of the important variations is the image-to-image translation GAN, which translates a given image in one domain to another. StarGAN2 and cycleGAN are such models that convert images between domains and uses a cyclic loss to establish a balance between the images originating from different domains. After training and generating images from the GAN models, HR-Net and MT-CNN are used to detect the eye centers. Next, these detected eye center coordinates are mapped to the original thermal images. The benefits of this approach are (i) it is designed to work efficiently in low light to no light environments, where visible images cannot be used; (ii) since the process is automated and highly accurate, no operator intervention is needed to manually annotate any landmarks on an input LWIR image.

The contributions of this work are:

- Geometric eye centers in the thermal face images are detected by synthesizing visible images from their thermal counterparts using image translation GAN models.
- Alignment loss is successfully utilized in training phase as an additional constraint to the StarGAN2 and cycleGAN models to preserve the alignment, which was introduced in Jesorsky et al., 2001 and is only used in test phase previously (Jesorsky et al., 2001; Mokalla & Bourlai, 2021b; Whitelam & Bourlai, 2015).
- The bench-mark eye center detection accuracy (Mokalla & Bourlai, 2021b) is improved by around 17%, 7%, and 3% when normalized error (e) $\leq$ 0.05, 0.10, and 0.25, respectively.
- We also improve the thermal same-spectral face recognition (FR) accuracy by 3-4%.

1.4 Effects of Demographics and Photometric Normalization on Image Translation GANs for Cross-Spectral Face Recognition

Similar to the eye center detection problem, cross-spectral face recognition does not have a straight-forward solution. This is because of the spectral gap between the visible and the thermal images. One approach to address this problem is to utilize the image translation models, as explained above, to synthesize visible images from their thermal counterparts. Then, any of the numerous visible band-based models can be utilized to perform same-spectral face recognition. This section analyzes the effects of various factors on GAN-based image translation models and thereby on the cross-spectral FR. The models explored in this study are pix2pix (Isola et al., 2017), and StarGAN2 (Choi et al., 2020). The goal of this study is to understand the effects of demographic information on the performance of these networks. The application of different photometric normalization techniques to face images of both the bands before synthesis to understand the positive and negative effects of these techniques is also explored.

The contributions of this study are explained below:

- To propose a synthesis based approach for thermal-to-visible face recognition.
- Qualitatively assess the effects of demographic information such as, ethnicity, and presence of beard on pix2pix, and
- Qualitatively assess the effects of different photometric normalization techniques on image synthesis and thereby on cross-spectral face recognition.

1.5 Involution GAN: Rethinking Architecture to Improve the Performance of Cross-Spectral Face Recognition

The next part of this dissertation focuses on developing models for cross-spectral face recognition for LWIR-visible and MWIR-visible matching. Both of these bands have been used in the FR applications, including same- and cross-spectral matching (Anghelone et al., 2022). However, developing an FR model in thermal or cross-spectral scenarios does not yield as competitive results in terms of accuracy as visible band ones (Benamara et al., 2022; Cheema et al., 2022). This can be attributed to the unavailability of sizeable datasets in thermal spectrum which arises from the cost of the sensors, and strict Institutional Review Board (IRB) regulations (used in the US institutions; there are similar boards in EU and other countries), and the spectral gap between the visible and thermal band images as shown in Figure 1.3. To address this problem, a solution, similar to the eye center detection , that leverages the available visible band FR models by synthesizing visible images from their thermal counterparts, is proposed. To synthesize the images, A new architecture of Generative Adversarial Networks (GANs) motivated by StarGAN2 (Choi et al., 2020) and involution (D. Li et al., 2021), which inverts the inherent principles of convolution, is implemented.

Specifically, the general basis of GANs is utilized, i.e., a generator and a discriminator, and include a style encoder, proposed in (Choi et al., 2020), and build each of these networks using involution operation (D. Li et al., 2021). Involution, that inverts the inherence of convolution for visual recognition tasks, is used as the basis for all the neural network blocks in the solution instead of the popular CNNs. This has proved to improve not only the accuracy, but also reduces the training time significantly. Additionally, we use a pre-trained ArcFace model (Deng et al., 2019) to extract feature embeddings from the original and the synthesized images and use the cosine distance between these embeddings as an additional loss term. The trained GAN model is then used to synthesize the visible images from the thermal images in the test data. Now the problem of cross-spectral FR is reduced to same-spectral FR for which the solution exists by means of the numerous visible band-based FR models (Deng et al., 2019; Schroff et al., 2015; H. Wang et al., 2018). This is depicted in Figure 1.5.

To address the problem of cross-spectral face recognition, a thermal-visible synthesis-based approach to leverage the available FR models is proposed. For this, the GAN architecture is redesigned using the involution kernels. Involution operation inverts the inherence of convolution i.e., convolutional kernels are spatial-agnostic and channel-specific, whereas involution kernels are channel-agnostic and spatial-specific. This property deprives the CNNs of the ability to adapt to diverse visual patterns at different spatial locations within an image. To overcome this, involution kernels are distinct in spatial extent but are shared across channels. The proposed GAN architecture includes a generator, a discriminator, and a style encoder and roughly follows the architectural guidelines stipulated in Choi et al., 2020. Extracted feature embeddings using ArcFace from the original and synthesized images are used to compute an additional identity loss which serves to preserve the distinguishable characteristics of individual subjects.

This solution presents the following key benefits: (i) Works efficiently to match a visible band gallery to a thermal band query; (ii) Performs well on a multitude of cross-spectral datasets at various distances, as indicated in section 4.4, where the results are presented using two different datasets. These are the Army Research Laboratory - Visible Thermal Face dataset (ARL-VTF) (Poster et al., 2021) and Multi-spectral Imagery Lab (MILAB) dataset (Bourlai et al., 2023) which consists of images from LWIR-Visible and MWIR-Visible bands, respectively.

The primary contributions of this paper can be summarized as follows:

- A solution for cross-spectral FR that eliminates the need to train or re-train an original deep learning model which requires images in the order of tens of thousands to millions is proposed.
- GAN architecture is redesigned using the involution operation to improve both, the speed and the accuracy.
- The SOTA verification Area Under the Curve (AUC) on ARL is increased by dataset by 4% and the equal error rate (EER) is decreased by 7%.
- The SOTA AUC increased by 17% and 10% and the EER decreased by 15% and 11% over MILAB-VTF(B) indoor and outdoor datasets, respectively.

The rest of the dissertation is organized as follows. Chapter 2 expands on the literature review related to all the problems addressed. Methodological approach followed in addressing each of the problems is illustrated in chapter 3. All the experiments performed and the results obtained are presented in chapter 4, and the work is concluded in chapter 5 along with providing insights into the future work.

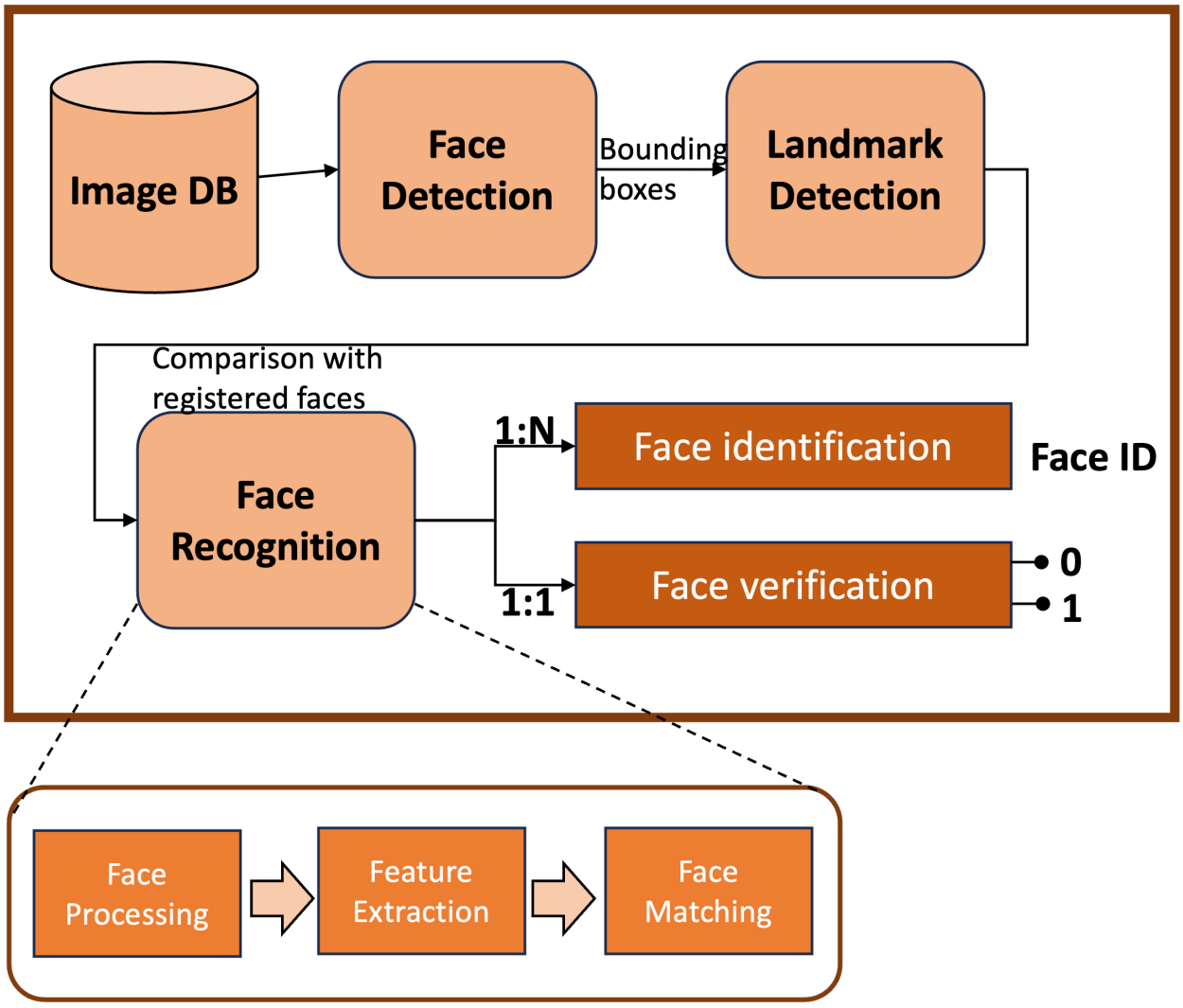

Figure 1.1: Building blocks of a typical face recognition system. Once the images are acquired, face detection is performed to obtain the bounding boxes, then, landmark detection is performed for face alignment. Features are extracted from this aligned face to be used for either identification or verification.

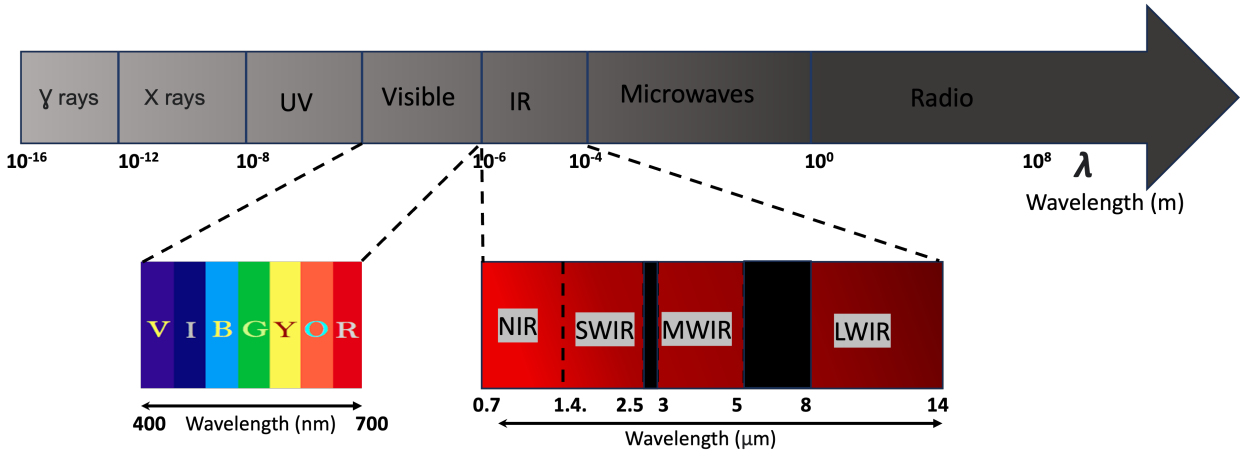

Figure 1.2: Various bands and their wavelengths within the electro-magnetic (EM) spectrum.

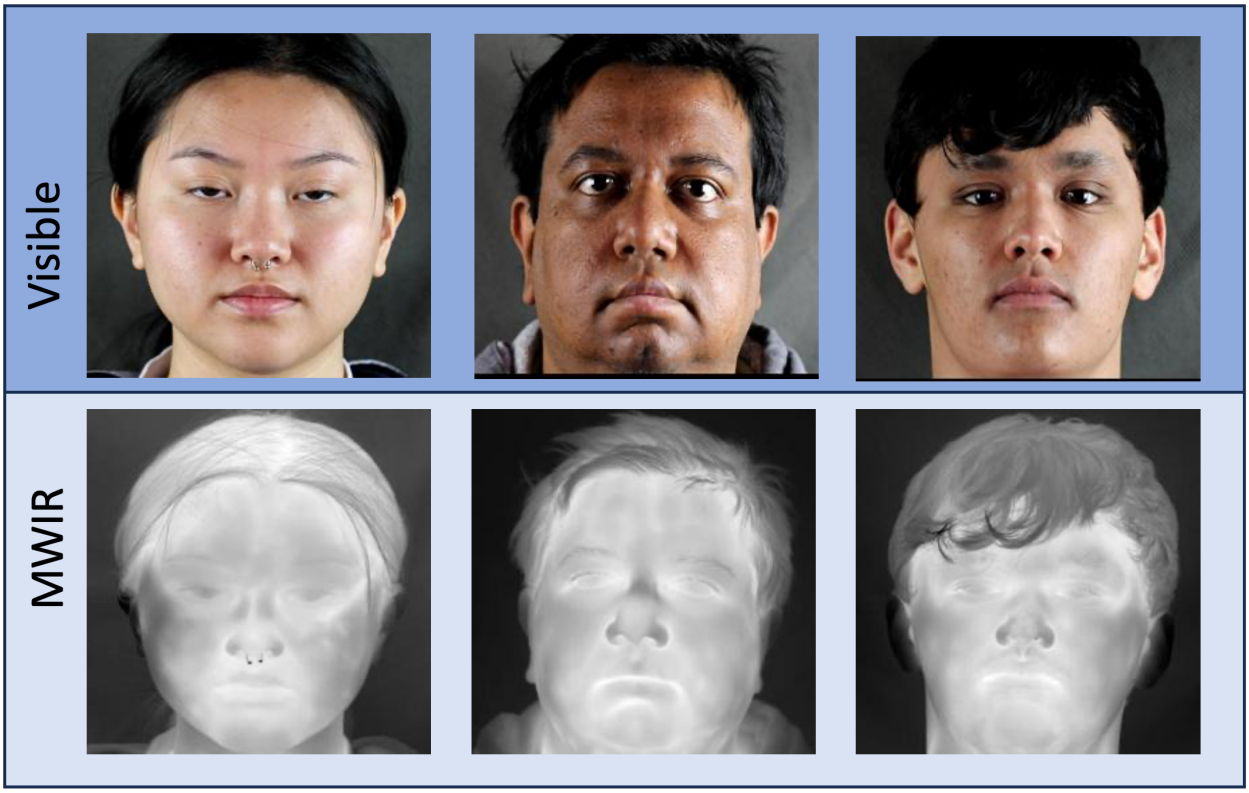

Figure 1.3: Example images from the MILAB-VTF(B) dataset.

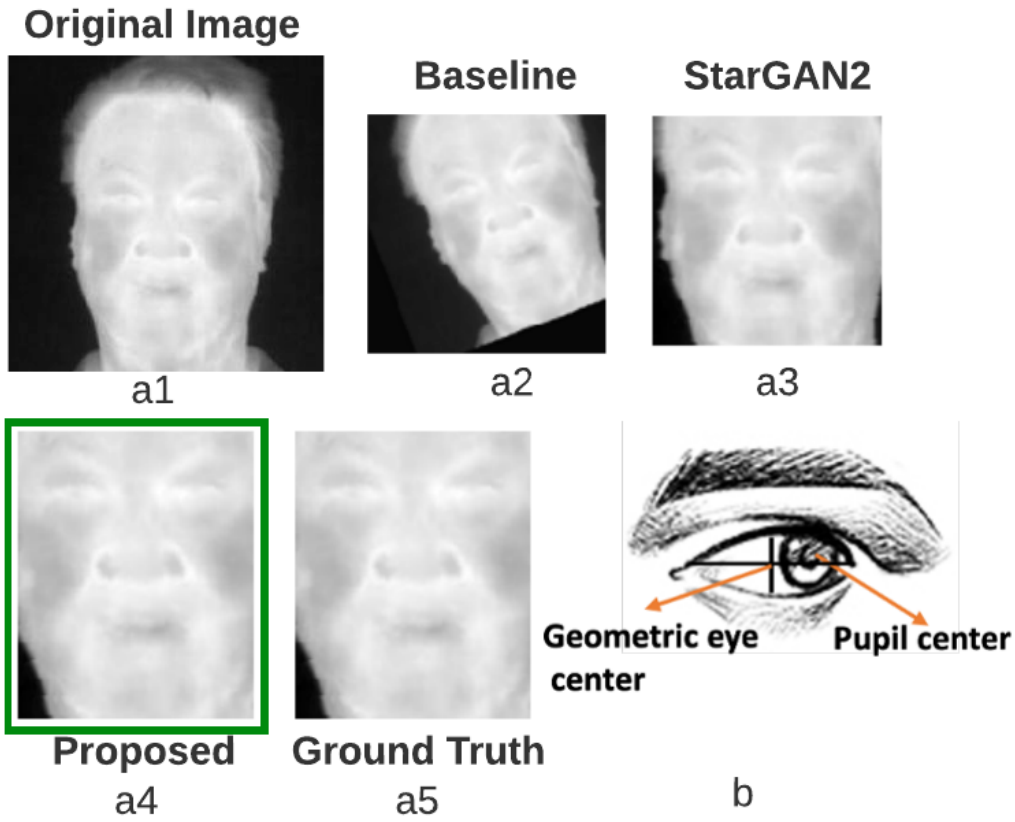

Figure 1.4: This figure shows the original image and geometrically normalized images using various methods. a1 - original image, a2 - Using the baseline method, a3 - Using StarGAN2, a4 - Using the proposed method, a5 - Using ground truth annotations. b. Figure explaining the difference between pupil and geometric eye centers. This work focuses on the latter for image alignment, while pupil center detection has other applications such as driver drowsiness detection, human-computer interaction etc.

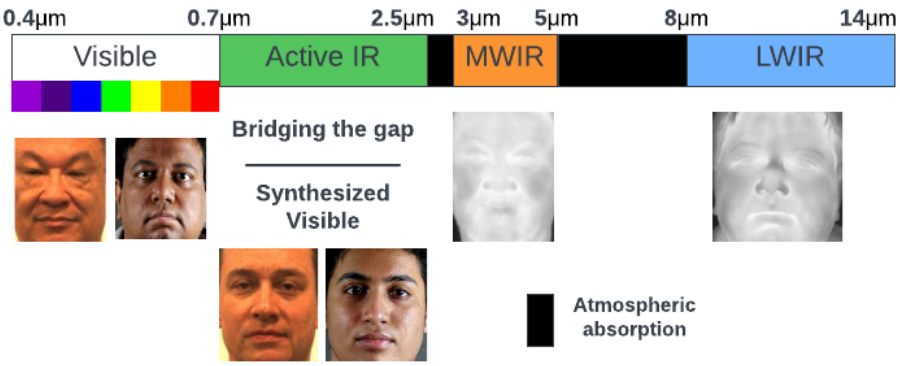

Figure 1.5: Figure showing the spectral gap bridged in this study. The gallery images are in the visible band of the electro-magnetic spectrum, and the query images are in the thermal spectrum (LWIR for ARL-VTF, and MWIR for MILAB-VTF(B) dataset). Face recognition is performed by synthesizing the visible images from the thermal images and utilizing the existing algorithms in the visible spectrum.

Chapter 2

Literature Review

2.1 Multi-spectral Face Dataset Collection

Table 2.1: The variable characteristics of each dataset are denoted as follows: (P)ose, (I)llumination, (E)xpression, (T)ime-lapse, (G)lasses, (O)cclusion, and (L)ocation. The subscript N is used to identify characteristics that occur due to natural outdoor conditions (i.e. sunlight, clouds, and wind). MILAB-VTF(B) uniquely captures high-resolution paired thermal and visible scenes outdoors at large distances. Importantly, the dataset is diverse with respect to ethnicity, age, and gender. This table is adapted from Poster et al., 2021.

Dataset	Modalities	Subjects	Variability
UND (Kevin & Bowyer, 2003)	LWIR, Visible	241	I, E, T
NVIE (S. Wang et al., 2010)	LWIR, Mono	215	I, E, G
ULFMT (Ghiass et al., 2018)	MWIR, Visible	238	P, E, T, G
ARL-MMFD (Hu et al., 2016)	P-L, LWIR, Visible	111	E
Tufts (Panetta et al., 2018)	NWIR, LWIR, Visible	100	P, E
ARL-VTF (Poster et al., 2021)	LWIR, Visible, Mono	395	P, E, G
MILAB-VTF(B)	MWIR, Visible	400	P, L, I_N, E_N, O_N

Tables 2.1 and 2.2 provides a comparison between MILAB-VTF(B) dataset and similar datasets that were created in different spectra. Specifically, the MILAB-VTF(B) raw dataset is characterized by the following key advantages compared to the other thermal-visible datasets:

1. It is the largest thermal-visible face dataset to-date in terms of number of subjects, images, and scenarios.

2. High resolution video and images at 1280×1024, double previous datasets.

3. Facial images and videos captured both under indoor and outdoor conditions.

Table 2.2: The variable characteristics of each dataset are denoted as follows: (P)ose, (I)llumination, (E)xpression, (T)ime-lapse, (G)lasses, (O)cclusion, and (L)ocation. The subscript N is used to identify characteristics that occur due to natural outdoor conditions (i.e. sunlight, clouds, and wind). MILAB-VTF(B) uniquely captures high-resolution paired thermal and visible scenes outdoors at large distances. Importantly, the dataset is diverse with respect to ethnicity, age, and gender. This table is adapted from Poster et al., 2021.

Dataset	IR Resolution (W × H)	Range (m)
UND (Kevin & Bowyer, 2003)	320 × 240	Unspecified
NVIE (S. Wang et al., 2010)	320 × 240	0.75
ULFMT (Ghiass et al., 2018)	640 × 512	1.0
ARL-MMFD (Hu et al., 2016)	640 × 480 (LW)	2.5, 5.0, 7.5
Tufts (Panetta et al., 2018)	336 × 256	1.5
ARL-VTF (Poster et al., 2021)	630 × 512	2.1
MILAB-VTF(B)	1280 × 1024	1.5, 100, 200, 300, 400

4. Natural face expression variations, collected mostly outdoors, where the weather dynamically impacted the facial expressions of participants.

5. Natural illumination and facial occlusion variations found in outdoor conditions resulting in shadows and facial hair obscuring part or sometimes the whole face region.

6. Data at five different stand-off distances, ranging from 1.5 meters controlled, and up to 400 meters (1312 ft) outdoors, in increments of 100 meters (100; 200; 300; 400 meters).

Finally, here are the common features, challenges, and benefits of the MILAB-VTF(B) dataset when compared to the second largest dataset, i.e. the ARL-VTF:

1. The indoor and outdoor raw face videos in MILAB-VTF(B) were recorded at the same time. Due to the complexity and the technical challenges of the data collection in an outdoor environment, as well as the limited time expected to deliver the dataset (6 weeks), the raw videos were not always synced. The videos are manually synced for the curated version of the dataset;

2. Both datasets were captured using commercially available thermal cameras;

3. Both datasets involve face data captured under variable facial expression and pose. In ARL-VTF the subjects count from 1-10 and the face images are captured under controlled conditions. In MILAB-VTF the facial expressions are natural and face images are captured under uncontrolled conditions.

4. The ARL-VTF offers a curated version that is publicly available (limited distribution) that includes automatically annotated facial landmarks on many face images samples. MILAB-VTF(B) curated version will also include facial landmarks in a smaller scale.

5. Finally, both datasets support algorithm development in a set of areas, including face/eye/ear detection, same- and cross-spectral face matching (Bourlai & Jafri, 2011), (Whitelam et al., 2010), (Mokalla & Bourlai, 2020), (Abaza & Bourlai, 2013), multi-modal fusion (Kakadiaris et al., 2005), domain adaptation, and cross-domain image synthesis (R. He et al., 2021). An example of synthesis approach for thermal-to-visible face verification is discussed in Isola et al., 2017.

2.2 Utilizing Alignment Loss to Advance Eye Center Detection and Face Recognition in the LWIR Band

There are several algorithms in the open literature that focus on eye center detection and localization in visible and thermal images using traditional and machine learning methods, a few algorithms use deep learning for the task in visible spectrum.

2.2.1 Deep Learning Based Methods

Very few algorithms available in the open literature for eye detection and localization use deep CNNs. One such work is (Nguyen et al., 2020), in which the authors use various convolutional layers and blocks such as a shrinking block to shrink the input image space by selecting an appropriate kernel size, an inception block, a convolutional block and a detection network. They use Region Proposal Network (RPN) loss function from Faster R-CNN (Ren et al., 2015). Fuhl, Santini, et al., 2016 proposed a dual CNN based pupil detection model for pupil detection that can be extended to eye center detection. The image is down-scaled and divided into overlapping sub-regions which are evaluated by the first CNN and the center of the sub-region that evokes the highest CNN response is fed into the second pipeline stage, which is called fine pupil position estimate.

Yu et al., 2018 proposed a deep CNN based method for eye detection in combination with Eye Variance Filter (EVF) and Support Vector Machines (SVM). While CNNs and SVMs are used as classifiers, EVF is constructed for eliminating most of the non-eye images to keep less candidate eye images. Ahmad et al., 2020 proposed a CNN based method for eye center localization where they train a convolutional neural network with Rectified Linear Unit (ReLU) and batch normalization layers using eye and non-eye images.

2.2.2 Eye Center Detection in the IR spectrum

Whitelam and Bourlai, 2015 proposed a template based method for eye detection in five different wavelengths of the Short Wave Infrared (SWIR) spectrum. They perform automatic face detection to reduce the search space for the eye detector, then further divide the face into four equal parts. Now, the left and

Table 2.3: Different Methods for Eye Localization in the Literature.

Literature	Method	CNNs	Band
Yu et al., 2018	CNNs and SVMs	**X**	Visible
Whitelam and Bourlai, 2015	Summation range filters	-	SWIR
Bourlai and Jafri, 2011	Internal projections	-	MWIR
Valenti and Gevers, 2011	Isophotes	-	Visible
Chaudhari and Kale, 2010	Gabor filters	-	Visible
G. Li et al., 2006	Brightness histogram	-	Visible
Mokalla and Bourlai, 2021b	Image synthesis	**X**	LWIR
Proposed	**synthesis+alignment loss**	**X**	**LWIR**

right eye centers are in the top left and right parts respectively. Then, summation range filters are used for accurate eye localization. Bourlai and Jafri, 2011 proposed another template based method for eye detection in the thermal spectrum where the images are first photometrically normalized using Contrast Limited Adaptive Histogram Equalization (CLAHE), then used an average from ten randomly selected subjects' images to generate eye and ocular region thermal templates and an internal projections method was used to detect eyebrows and eyelashes.

Whitelam et al., 2010 proposed a multi-spectral eye detection model composed of six modules: 1. Face width estimation, 2. Eye template generation, 3. Pupil diameter estimation for each generated eye template, 4. Face detection, 5. Usage of the face boundaries to assist in placing the generated eye templates, and 6. Lastly, within each detected eye, the pixel with the lowest intensity is found out and the location of this pixel is registered as the detected eye center.

2.2.3 Eye Center Detection in the Visible Spectrum

Valenti and Gevers, 2011 proposed a novel eye location detection approach based on the observation that eyes are characterized by radially symmetric brightness patterns. They use isophotes to infer the center of semi-circular patterns and gain invariance to in-plane rotation and linear lighting changes. They also introduce a center voting mechanism based on gradient slope in the isophote network to increase and weigh important votes to reinforce the center estimates and this is integrated into a scale space framework to find the most stable results. G. Li et al., 2006 proposed a face normalization algorithm by locating the position of eyeballs mainly based on the brightness histogram using pattern matching. Once the eyeballs are detected, the eye centers are detected and the face coordinate is obtained by the line segment joining the two eye centers.

Chaudhari and Kale, 2010 proposed two approaches for geometric face normalization, namely holistic and feature based. In the holistic approach, the 15% parts with the smallest gray values are chosen as the eyeball candidate regions and a segment threshold is used to determine the eye centers. Then the mouth center is localized followed by face normalization by rotating the image so that the line segment joining the eye centers is horizontal. A feature geometric face recognition approach uses a modified bunch graph

matching, which is a combination of graph isomorphism (features such as eye centers, nose tip, left and right corners of mouth are manually selected, and these five features are connected to each other to form a face graph) and 80 Gabor filters are generated and convolved with each of the features. After creating the face graphs for two images, the similarity is computed by comparing Gabor jets using the Euclidean distance.

Timm and Barth, 2011 proposed an eye detection model based on gradients. In their paper, they analyze the vector field of image gradients that exploits the flow field character that arises due to the strong contrast between iris and sclera. They use the orientation of each gradient vector to draw a line through the whole image and they increase an accumulator bin each time one such line passes through it. Thus, the accumulator bin, where most of the lines intersect, thus represents the eye center. In addition to this, they derive a relationship between a possible center and the orientations of all the image gradients. They use the vector field of gradients by computing the dot products between normalized displacement vectors and the gradient vectors.

Shafi and Chung, 2009 proposed a feature-based method for eye localization in visible band facial images, which includes extracting connected regions from facial regions using color, edge density and illumination cues separately. Some of the regions are then removed by applying rules that are based on the general geometry and shape of eyes and the remaining connected regions are combined in a systematic way to enhance the identification of the candidate regions for the eyes. The geometry and shape rules are once again applied to further remove any false eye regions. Park et al., 2007 proposed another feature based method in which facial regions of the input image are extracted automatically using AdaBoost, in order to minimize the illumination effects of various lighting conditions, the illumination normalization is carried out using local information of image. Then, they estimate the candidates of eyes based on texture information, and finally the eye centers are localized by geometrical information of the face.

W. Huang and Mariani, 2000 proposed a face detection and facial feature localization algorithm using multi-scale filters to obtain the pre-attentive features of objects to locate the face and facial features such as eyes, nose and mouth. A structural model is used to characterize the geometric pattern of facial components, and texture and feature models are used to verify the face candidates detected earlier. Then an algorithm to detect eye centers is proposed using contour and region information, and these eye center locations are used to normalize faces for recognition.

Xingming and Huangyuan, 2006 proposed an illumination independent algorithm for automatic eye localization in visible images using an illumination normalization method to overcome varying lighting conditions, then they use a pose independent AdaBoost method to extract face feature candidates. Finally a heuristic rule is used to filter non-eye candidates and an SVM is employed to classify the regions as eye/non-eye. Asteriadis et al., 2006 proposed an eye localization method where the faces are detected first and the edge map is extracted, then a vector is assigned to every pixel pointing to the closest edge pixel. The length and slope information of these vectors is then used to detect and localize the eyes. Table 2.4 shows that the proposed method is different from the previously proposed approaches.

2.3 Same-Spectral Face Recognition

There are some recent visible-to-visible face recognition algorithms available in the open literature that use deep CNNs (Convolutional Neural Networks). Most, if not all, of these models, use the distance between Euclidean or another form of embeddings as a measure of similarity between faces. FaceNet (Schroff et al., 2015) is one of the state-of-the-art face recognition algorithms available for visible-to-visible face recognition. It uses a triplet loss function which is obtained by calculating the L2 distance between the Euclidean embeddings of faces such that the distances represent the face similarity i.e., faces of the same person have smaller distances and faces of distinct people have large distances. Y. Li, 2019 presents an implementation of triplet loss on a face recognition task and conducts several experiments to analyze the factors that influence the training of triplet loss. They use triplet pairs (one anchor image – one positive image, the same anchor image – one negative image). The samples are then mapped into a feature vector through deep CNNs such as Resnet (K. He et al., 2016) or MobileNet (Howard et al., 2017). One major drawback of this method is that it requires the mining of triplets to train the model. Also, the angular margin is preferred to the Euclidean margin because the cosine of the angle has an intrinsic consistency with softmax. To overcome this problem, W. Liu et al., 2017 propose SphereFace, in which angular softmax (A-Softmax) loss is introduced that incorporates an angular margin. A-Softmax loss learns discriminating features that span on a hypersphere manifold, which intrinsically matches the prior that faces also lie on a manifold. In addition to that, they introduce a parameter that quantitatively controls the size of angular margin and derives lower bounds on this margin such that A-Softmax loss can approximate that minimal inter-class distance is larger than the maximal intra-class distance. This decision boundary parameter is defined over the angular space.

CosFace (H. Wang et al., 2018) uses Large Margin Cosine Loss (LMCL) that takes the normalized features as input to learn highly discriminating features by maximizing the inter-class cosine margin. This loss defines the decision margin in the cosine space and not in the angular space as in (W. Liu et al., 2017). ArcFace (Deng et al., 2019) utilizes the arc-cosine function to calculate the angle between the current feature and the target weight since the dot product between the DCNN (Deep Convolutional Neural Network) feature and the last fully connected layer is equal to the cosine distance after feature and weight normalization. Then, an additive angular margin is added to the target angle and the target logit is obtained back from the cosine function. This has an advantage over the other angular margin losses as it directly optimizes the geodesic margin which is the exact correspondence between the angle and the arc in the normalized hypersphere.

The above margin-based face recognition methods are susceptible to the label noise in the training data and thus require human effort to clean the datasets. To address this problem, Deng et al., 2020 propose the sub-center ArcFace, where the intra-class constraint forces all samples close to the corresponding positive centers by introducing sub-centers. This avoids the possibility of a noisy image not belonging to the corresponding positive class. Instead, several sub-centers are designed within the same class and the training sample only needs to be close to any of these sub-centers. This encourages one dominant

sub-class that contains the majority of the clean faces and multiple non-dominant sub-classes that include hard or noisy faces resulting in a model that is robust to noise.

Y. Sun et al., 2020 propose the Circle Loss, which penalizes various similarity scores differently i.e. if a similarity score deviates far from the optimum, it receives a significant penalty. To this end, they use two different weighting parameters for the intra-class and inter-class distances, allowing them to learn at different paces. This leads to a unified loss function that learns with class-level labels (softmax cross-entropy loss function) and pair-wise labels (triplet loss etc).

2.4 Cross-Spectral Face Recognition

To match images from two different modalities, one of the two methods are popularly used. First method is to project all the images to a common subspace, and the second is to synthesize images.

2.4.1 Feature based Methods

Common subspace projection (CSP) also known as Hashing is one of the well-known approaches for cross-spectral FR. CSP operates by projecting features extracted from different domains onto a common subspace and these projected features are used for direct matching since they are in the identical domain. Some of the more traditional hashing solutions for inter-domain retrieval are canonical correlational analysis (CCA), bilinear model (BLM), and partial least squares (PLS). The disadvantage with these methods is that they only consider inter-domain similarity measures on common subspace. To address this problem, H. Wang et al., 2021 proposed subspace projection hashing (SPH), in which face features are extracted as 512-dimensional vectors. These vectors are used to generate hashing matrices, which along with the features are projected onto the common subspace to generate hashed codes, which are used for face matching. Additionally, a new loss function that optimizes both inter-domain and intra-domain similarities is introduced.

Hu et al., 2015 proposed a thermal-to-visible FR method using partial least squares, which has a pre-processing stage, a feature extractor, and a partial least squares-based modeling. The pre-processing stage consists of four components: median filtering of dead pixels, geometric normalization through an affine transform to a common set of canonical coordinates for alignment, Difference-of-Gaussian (DoG) filtering, which is a common technique to remove illumination variations for visible FR, and constant normalization, which further enhances the edge information of the DoG filtered thermal and visible images. Then, features are extracted using Histogram of Oriented Gradients (HOG) method (Dalal & Triggs, 2005). Finally, these extracted features are matched using partial lease squares method (Wold, 1966).

Chen and Ross, 2016 proposed a thermal-to-visible framework, which uses multiple sets of subspaces generated by sampling patches from visible and thermal face images. These patches are represented by either a Pyramid Scale Invariant Feature Transform (PSIFT) or Histograms of Principal Oriented Gradients (HPOG). Then, they use a cascaded subspace learning process consisting of whitening transformation,

Table 2.4: Different Methods for thermal-to-visible FR in the Literature.

Literature	Method	Synthesis	Involution
H. Wang et al., 2021	SPH	-	-
Hu et al., 2015	PLS	-	-
Chen and Ross, 2016	PSIFT and HPOG	-	-
Di et al., 2021	AP-GAN	**X**	-
Chen and Ross, 2019	Multi-scale GAN	**X**	-
Benamara et al., 2022	GAN+SSIM	**X**	-
Mallat et al., 2019	CRN+contextual loss	**X**	-
Proposed	Involution GAN +identity loss	**X**	**X**

factor analysis, and common discriminant analysis is used to construct multiple common subspaces. At the end, Nearest Neighbor (NN) is used to compare the feature vectors.

Riggan et al., 2018 proposed an optimal feature learning and discriminative framework for polarimetric thermal to visible face recognition which consists of a deep perceptual mapping (DPM) for direct regression and coupled neural network (CpNN) for indirect regression. For DPM, HOG or Scale-Invariant Feature Transform (SIFT) features are extracted from the visible and IR domains and Principal Component Analysis (PCA) is used for dimensionality reduction, which in turn reduces the number of trainable parameters. The CpNN assumes one fixed mapping to ensure that both DPM and CpNN are performing a one-way regression i.e., thermal-to-visible. Finally, a discriminative regression method called PLS is used for recognition.

Osia and Bourlai, 2017 formulated an image synthesis framework and post-synthesis restoration methodology, to improve cross spectral FR accuracy. They explored cohort-specific matching (per gender) instead of blind-based matching (when all images in the gallery are matched against all in the probe set). Narang and Bourlai, 2015, Bourlai et al., 2012, and Osia and Bourlai, 2014 are other examples of similar approaches for FR outside visible.

2.4.2 Synthesis based Methods

The feature-based methods discussed above are typically two-step methods, where images from the two domains are represented and projected onto a common subspace and these projections are used for matching. Synthesis is a one-step method where visible images are generated from their corresponding thermal images. Since the resultant images from these methods are in the visible spectrum, any of the existing visible band-based FR algorithms (Deng et al., 2019; Schroff et al., 2015; H. Wang et al., 2018) can be used

for matching. As a result, GANs are currently the most popular image synthesis models for cross-spectral FR.

Most, if not all, of the GAN based thermal-to-visible synthesis models use pix2pix (Isola et al., 2017) or CycleGAN (Zhu et al., 2017) as the base model and add different loss functions and feedback networks to improve the cross-spectral face verification accuracy. One of the first works that follow this methodology are Semantic-Guided GAN (SG-GAN) (Chen & Ross, 2019), in which a face parsing network (S. Liu et al., 2015) is used to extract semantic priors from the generated and the target visible images and calculate a semantic loss. Additionally, VGGFace is used to calculate identity loss and VGG-19 pre-trained network is used to extract features and a perceptual loss is calculated. Attribute Preserved GAN (AP-GAN) (Di et al., 2021) is another such work that uses an attribute predictor which is a fine-tuned VGGFace network and is trained separately from multi-AP-GAN for polarimetric thermal to visible image synthesis. Similar to Chen and Ross, 2019, identity loss and perceptual losses are used along with the GAN and the multi-scale attribute losses.

Auto-encoders are another class of multi-layer perceptrons that are sometimes used along with GANs, as proposed in Patel and Upla, 2021, where the proposed architecture includes two auto-encoders, two generators, and two discriminator modules. One generator is used to generate an image from the reconstructed stream and the other is trained to generate an image from the translated stream. Zhang et al., 2017 proposed a GAN approach that uses the aforementioned identity and perceptual losses and a guidance sub-network (Xie & Tu, 2015) at the end of the visible feature extraction part to guarantee the reconstructability of the encoded features and to make sure that the learned features contain semantic information.

Below are some of the more recent works that use GANs to synthesize visible images from thermal or polarimetric thermal images. Cycle Synthesized Attention GAN (CSA-GAN) (Yadav et al., 2022) uses attention guidance and cyclic synthesis objective to reduce the learning space, thereby, leaning towards finding an optimal solution. Additionally, structural similarity index measure (SSIM) and multi-scale SSIM (MS-SSIM) losses are used to improve the training capabilities of the attention network.

Benamara et al., 2022 proposed a heterogeneous FR based on CycleGAN to synthesize visible images from LWIR face images by incorporating SSIM. They also proposed a multi-sensor detector based on the recent YOLO v2 architecture for face detection in visible as well as LWIR imagery. Cheema et al., 2022 proposed Cross-Modality Discriminator Network (CMDN) for heterogeneous FR to learn deep feature relations for cross-domain face matching while simultaneously extracting modality-independent embedding vectors for face images. The proposed CMDN parameters are optimized using a novel Unit-Class loss which is a weighted sum of the triplet loss (Schroff et al., 2015) and the class mean triplet loss (where triplet loss is calculated between the sample and the mean of the classes). Mokalla and Bourlai, 2021a conducted several experiments to understand the impact of various photometric normalization techniques and demographics on the performance of GAN models for face verification.

Mallat et al., 2019 proposed the usage of cascaded refinement networks (CRN) coupled with contextual loss to synthesize high quality-colored visible images from thermal acquisitions. CRN considers multi-scale information and is based on training a limited number of parameters, and employing the

contextual loss makes this method scale and rotation invariant. Peri et al., 2021 proposed an end-to-end thermal-to-visible synthesis method that includes face detection, keypoint regression and matching. To train a synthesis model, they incorporated an identity loss by extending the perceptual loss with an additional constraint such that all the generated images of a particular class must cluster together. This loss is used in addition to the general-purpose domain-adaptation loss and pixel-wise loss. Some other works that use synthesis methodology for other tasks are (Mokalla & Bourlai, 2023), (Mokalla & Bourlai, 2021b).

Chapter 3

Methodology

This chapter explains the methodological approach followed to address each of the problems discussed in Chapter 1 in great detail. Each of the components used are described in a thorough manner.

3.1 Utilizing Alignment Loss to Advance Eye Center Detection and Face Recognition in the LWIR Band

First, baseline experiments are performed by training the Multi-task Cascaded CNN (MTCNN) using the original thermal images for face and landmark detection. Then, MTCNN, pretrained on visible data is incorporated into the training stage of StarGAN2 and CycleGAN to get the annotations out of the synthesized images. The normalized error between these detected eye centers and the ground-truth annotations is calculated, and is used as an additional loss term. Results obtained show that the inclusion of this new loss term improves the eye center detection accuracy and thereby the face recognition accuracy. During test stage, the trained model is used to generate visible face images from their corresponding thermal images. Then, pre-trained landmark detection models, namely, MT-CNN and HR-Net, are used to detect the eye centers in the synthesized images. These eye center coordinates are mapped on to the original thermal images to geometrically normalize and perform same spectral face recognition. The methodology is illustrated in Figure 3.1.

3.1.1 MT-CNN

Multi-Task Cascaded Convolutional Neural Network (MT-CNN) Xiang and Zhu, 2017 is a lightweight framework that integrates the two tasks, face detection and face alignment using unified cascaded CNNs by multi-task learning. Given an image pyramid, it is resized to different scales to build an image pyramid, which is the input to the following three-stage framework:

Stage 1: The first stage is called proposal Network (P-Net), which is a fully convolutional network. This is used to obtain candidate windows and their bounding box regression vectors. Then, the esti-

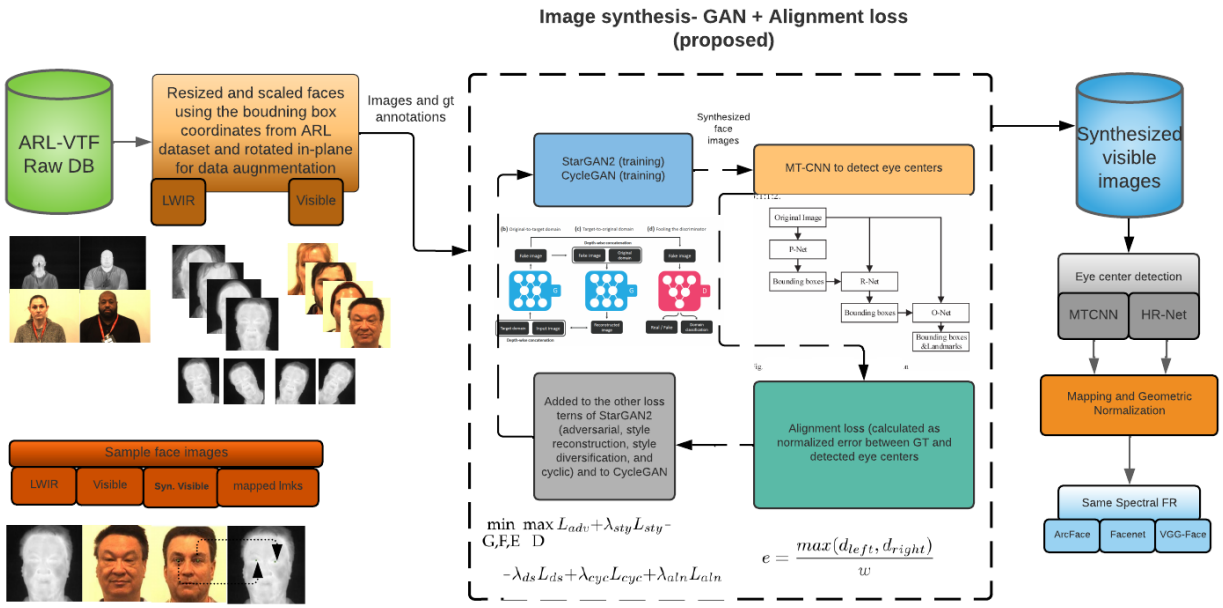

Figure 3.1: The methodological approach followed to address the problem of eye center detection in the LWIR spectrum through image synthesis. Raw images are cropped and resized to satisfy the GAN models' training and test requirements. Data augmentation is performed by rotating the train and test images in-plane. These resized images are then used to train GAN models with an additional loss term (alignment loss). This requires including MT-CNN in the pipeline to detect the eye centers in the synthesized images during training, and calculating the normalized error.

mated bounding box regression vectors are used to calibrate the candidates. After that, non-maximum suppression (NMS) is employed to merge highly overlapped candidates.

Stage 2: At this stage, all the candidates are fed to another CNN, called Refine Network (R-Net), which further rejects a large number of candidates, and performs calibration with bounding box regression, NMS candidate merge.

Stage 3: This stage is similar to the second stage, but the aim here is to describe the face in more detail. In particular, the network outputs five facial landmarks' positions.

Three tasks are leveraged to train the CNN detectors: face/non-face classification, bounding box regression, and facial landmark localization. All the three stages are presented in Figure 3.2.

1. Face classification: The learning objective is formulated as a two-class classification problem. For each sample x_i, cross-entropy loss is used:

$$L_i^{det} = -(y_i^{det} log(p_i) + (1 - y_i^{det})(1 - log(p_i)) \tag{3.1}$$

where p_i is the probability produced by the network that indicates a sample being a face. The notation $y_i^{det} \in \{0, 1\}$ represents ground-truth label.

2. Bounding box regression: The offset between each candidate window and the nearest ground truth (the bounding box's left top coordinate, height, and width) is predicted. Then, the learning objective is formulated as a regression problem, and Euclidean loss is employed for each sample x_i:

$$L_i^{box} = \|\hat{y}_i^{box} - y_i^{box}\|_2^2 \tag{3.2}$$

where $\hat{y}_i^{box}$ is the regression target obtained form the network and y_i^{box} is the ground truth coordinate. There are four coordinates, including the top-left coordinates, height and width, and thus, $y_i^{box} \in R^4$.

3. Facial landmark localization: Similar to the bounding box regression task, facial landmark detection is formulated as a regression problem and the Euclidean loss is minimized:

$$L_i^{lmk} = \|\hat{y}_i^{lmk} - y_i^{lmk}\|_2^2 \tag{3.3}$$

where $\hat{y}_i^{lmk}$ is the facial landmark's coordinate obtained from the network and y_i^{lmk} is the ground truth coordinate. There are five facial landmarks, namely, left eye, right eye, nose, left mouth corner, and right mouth corner, therefore $y_i^{lmk} \in R^{10}$.

3.1.2 Generative Adversarial Networks

GANs are first proposed by Goodfellow et al., 2014. It has two multi-layer perceptrons, called the generator and the discriminator. The generator takes noise as input and outputs images. The discriminator takes the original and the generated images as input and classifies these images as either real or fake. The generator is trained to be able to generate images that are indistinguishable from the original images, and the discriminator is trained to correctly classify the images as real or fake on each occasion. The generator network can be thought of as analogous to a team of counterfeiters, trying to produce fake currency and use it without detection, and the discriminator is analogous to the police trying to detect the counterfeit currency. Competition in this game drives both teams to improve their methods until the counterfeits are indistinguishable from the genuine currency. In other words, discriminator, D, and generator, G play a mini-max game with the following objective:

$$\min_G \max_D V(D, G) \quad = \quad E_{x \sim p_{\text{data}}(x)}[\log D(x)] \quad + \quad E_{z \sim p_z(z)}[\log(1 \quad - \quad D(G(z)))] \tag{3.4}$$

where x denotes the real data samples drawn from the true data distribution $p_{\text{data}}(x)$, and z represents the noise vector drawn from the prior distribution $p_z(z)$. The objective function is composed of two terms. The first term, $Ex \sim p\text{data}(x)[\log D(x)]$, measures the discriminator's ability to correctly classify real data samples as real. The discriminator aims to maximize this term by assigning high probabilities

to real samples. The second term, $E_{z\sim p_z(z)}[\log(1 - D(G(z)))]$, measures the discriminator's ability to correctly classify fake data samples generated by the generator as fake. The generator aims to minimize this term by generating samples that can fool the discriminator. The overall objective of the GAN is to find the optimal generator G that minimizes the discriminator's ability to distinguish between real and fake samples while simultaneously maximizing the discriminator's accuracy in classification. The basic functionality of GAN is shown in Figure 3.3.

3.1.3 CycleGAN

After GANs were introduced, many variations of the original concept have been proposed, one being the conditional GAN (Mirza & Osindero, 2014), which has a generator and a discriminator conditioned by a label. Based on conditional GANs, image-to-image translation GANs have been proposed to convert a given image from one modality to another. However, many image-to-image translation GANs (Isola et al., 2017) require paired images. In case of thermal-to-visible face image synthesis, this requires geometrically normalizing the faces, which is the end goal of this research study. Therefore, an image translation model that does not require paired images, therefore does not require eye center annotations needed to be used.

CycleGAN is such an algorithm that learns to transfer between domains without paired input-output examples, therefore incorporating supervision at the level of domains. Assume there are one set of images in domain X and a different set in domain Y. Mathematically, if there are two translators, $G : X \rightarrow Y$ and $F : Y \rightarrow X$, then G and F should be inverses of each other. The cyclic loss used to train cycleGAN encourages $F(G(x)) \approx x$ and $G(F(y)) \approx y$. Combining this loss with adversarial losses on domains X and Y yields the full objective for unpaired image-to-image translation of this GAN.

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) = {} & E_{y\sim p_{\text{data}}(y)}[\log D_Y(y)] \\ & + E_{x\sim p_{\text{data}}(x)}[\log(1 - D_Y(G(x)))]\end{aligned} \tag{3.5}$$

where G tries to generate images $G(x)$ that look similar to images form domain Y, while D_Y aims to distinguish between translated samples $G(x)$ and real samples y. G aims to minimize this objective against the adversary D that tries to maximize it i.e., $min_G max_{D_X} \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y)$. In a similar fashion, for the mapping $F : Y \rightarrow X$ and its discriminator D_X:

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(F, D_X, Y, X) = {} & E_{x\sim p_{\text{data}}(x)}[\log D_X(x)] \\ & + E_{y\sim p_{\text{data}}(y)}[\log(1 - D_X(G(y)))]\end{aligned} \tag{3.6}$$

and similar to the adversarial loss in 3.5, the objective becomes $min_G max_{D_Y} \mathcal{L}_{\text{GAN}}(F, D_X, Y, X)$.

The cycle consistency loss to preserve the cyclic nature between the two domains is defined as:

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G, F) = {} & E_{x\sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] \\ & + E_{y\sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1]\end{aligned} \tag{3.7}$$

The cycle consistency is explained in Figure 3.4.

3.1.4 StarGAN2

StarGAN2 (Choi et al., 2018) is another such variation of GAN which is a scalable approach that can generate diverse images across multiple domains. It introduces two modules to replace StarGAN's domain, a mapping network, and a style encoder. The mapping network learns to transform random Gaussian noise into a style code, while the encoder learns to extract the style code from a given reference image. When using multiple domains, both modules have multiple output branches each providing a style code for a specific domain. The generator learns to successfully synthesize diverse images over multiple domains utilizing these style codes. The overall network architecture of StarGAN2 consists of four modules i.e. the generator, the mapping network, the style encoder, and the discriminator.

The generator module translates a given input image into an output, reflecting a specific style code s, provided either by the mapping network or by the style encoder E. The mapping network consists of a multi-layer perceptron with multiple output branches, to provide style codes for all the available domains. The style encoder extracts the style from an input image and produces diverse style codes using different reference images. This allows the generator to synthesize an output image reflecting the style of a reference image. Here, the source image contains the object, and the reference image contains the style. Therefore, in this work, thermal face images are the source, and visible images are the reference. The last module, a multi-task discriminator, consists of multiple output branches. Each branch learns a binary classification, determining whether an image produced by the generator is either real or fake. All the modules and their functionality is shown in Figure 3.5.

Once the model is trained, one of the two synthesis approaches can be used to generate images in the testing phase, which are latent-guided synthesis and reference-guided synthesis. In latent-guided synthesis, images are generated by taking a latent code from the trained model at random i.e., the generated image has the same object as that of the input image and the style from any of the images in the other domain. In the reference-guided synthesis, each input or source image needs a reference image in the respective domains and the generated image uses the object from the source image and the style from a user selected reference image. Since this approach uses face images of the same subjects from the visible and thermal bands, reference-guided synthesis is used.

The adversarial loss of StarGAN2 is defined as:

$$\begin{aligned}\mathcal{L}_{\text{adv}} = E_{x,y}[logD_y(x)] \\ + E_{x,\tilde{y},z}[log(1 - D_{\tilde{y}}(G(x,\tilde{s})))]\end{aligned} \tag{3.8}$$

where $D_y(.)$ denotes the output of D corresponding to the domain y. The mapping network F learns to provide a style code $\tilde{s}$ that is likely in the target domain $\tilde{y}$, and G learns to utilize $\tilde{s}$ and generate an image $G(x, \tilde{s})$ that is indistinguishable from the real images of the domain $\tilde{y}$. Then, in order to enforce the generator G to utilize the style code $\tilde{s}$ when generating the image $G(x, \tilde{s})$, a style reconstruction loss is employed:

$$\mathcal{L}_{\text{sty}} = E_{x,\tilde{y},z}[\|\tilde{s} - E_{\tilde{y}}(G(x,\tilde{s}))\|_1] \tag{3.9}$$

At test time, the learned Encoder E allows G to transform an input image, reflecting the style of a reference image. Finally, the cyclic loss used to properly preserve the domain-invariant characteristics of its input image **x** is given below:

$$\mathcal{L}_{\text{cyc}} = E_{x,y,\tilde{y},z}[\|x - G(G(x, \tilde{s}), \tilde{s})\|_1] \tag{3.10}$$

where $\tilde{s} = E_y(x)$ is the estimated style code of the input image **x**, and y is the original domain of **x**. By encouraging the generator G to reconstruct the input image **x** with the estimated style code $\tilde{s}$, G learns to preserve the original characteristics of **x** while changing its style faithfully.

3.1.5 Alignment Loss

The alignment loss added to the overall loss function to train StarGAN2 and CycleGAN is the normalized error from Jesorsky et al., 2001, given by

$$e = \frac{max(d_{left}, d_{right})}{w} \tag{3.11}$$

where d_{left} and d_{right} are the Euclidean distances between the detected and manually annotated left and right eye centers respectively, and *w* is the Euclidean distance between the ground-truth left and right eye centers. This can be defined as the weakest eye estimation (since the maximum of the distances between the true and the estimated eye centers is taken) divided by the distance between the true eye centers. Normalizing by dividing the worst estimation with the distance between the true eye centers makes it independent of scale of the face in the image and the image size. This is illustrated in detail in Figure 3.6.

3.1.6 Training Objective

The original StarGAN2 training objective is shown below:

$$\min_{\text{G,F,E}} \max_{\text{D}} L_{adv} + \lambda_{sty} L_{sty} - \lambda_{ds} L_{ds} + \lambda_{cyc} L_{cyc} \tag{3.12}$$

where L_{adv} is the adversarial loss, L_{sty} is the style reconstruction loss, L_{ds} is the style diversification loss, and L_{cyc} is the cyclic loss (Zhu et al., 2017), and λ_{sty}, λ_{ds}, and λ_{cyc} are the hyper-parameters for each term. Each of these terms are explained in great detail in equations 3.8, 3.9, and 3.10.

CycleGAN can be viewed as training two "autoencoders" where one autoencoder $F \circ G : X \rightarrow X$ is learnt jointly with another $G \circ F : Y \rightarrow Y$. However, these autoencoders each have special internal structures - they map an image to itself via an intermediate representation that is a translation of one image to another The original CycleGAN training objective is given below:

$$G^*F^* = arg \min_{\text{G,F}} \max_{\text{D}_X,\, D_Y} [L_{GAN}(G, D_Y, X, Y) + L_{GAN}(F, D_X, Y, X) + \lambda_{cyc} L_{cyc}(G, F)] \quad (3.13)$$

where D_X and D_Y are the two adversarial discriminators that aim to distinguish between images x and translated images F(y), and between images y and translated images G(x) respectively. Each of these terms are explained in equations 3.5, 3.6, 3.7.

Alignment loss is incorporated into the objectives in 3.23 and 3.13. The new training objective for StarGAN2 is given below:

$$\min_{\text{G,F,E}} \max_{\text{D}} \quad L_{adv} + \lambda_{sty} L_{sty} - \lambda_{ds} L_{ds} + \lambda_{cyc} L_{cyc} + \lambda_{aln} L_{aln} \quad (3.14)$$

where L_{aln} is the alignment loss given in Equation 3.11, and λ_{aln} is the hyperparameter.

Similarly, the new training objective for CycleGAN is below:

$$\begin{aligned} G^*F^* = arg \min_{\text{G,F}} \max_{\text{D}_X,\, D_Y} & [L_{GAN}(G, D_Y, X, Y) \\ & + L_{GAN}(F, D_X, Y, X) \\ & + \lambda_{cyc} L_{cyc}(G, F) + \lambda_{aln} L_{aln}] \end{aligned} \quad (3.15)$$

3.1.7 Eye Center Detection

Once the GAN models, supported by the proposed alignment loss, are trained and visible images are generated from the original thermal images, HR-Net (K. Sun et al., 2019) and MT-CNN (Xiang & Zhu, 2017) are used to detect the geometric eye centers. Then, these eye center coordinates are mapped to the associated original thermal face images as shown in Figure 3.7. The accuracy of the eye detection is determined using the normalization error in Eq 3.11. Then, a set of same-spectral face recognition experiments are performed using Facenet (Schroff et al., 2015), ArcFace (Deng et al., 2019), and VGG-Face (Parkhi et al., 2015). Training details and obtained results are presented in Section 4.2.

3.2 Effects of Demographics and Photometric Normalization on Image Translation GANs for Cross-Spectral Face Recognition

This section explains the methodology of the proposed approach in detail. First, the images are cropped and resized for StarGAN2, and are geometrically normalized for pix2pix. Then, the models are trained using these images. Next, these trained models are tested using the test images, and the synthesized images

from the test set are used for face verification using Facenet (Schroff et al., 2015). An overview of the proposed approach is shown in Figure 3.8.

3.2.1 Pix2pix

Unlike the unpaired GANs explained in section 3.1, Pix2pix is an image-to-image translation model that utilizes paired images. It is based on the property of conditional GANs, where the input to the generator is an image along with the noise vector, and the generator and the discriminator networks are conditioned on another image captured from a different modality. Given the generator G, and the discriminator D, input image x, noise vector z, and the conditional or the reference image y, the objective function of a pix2pix network is given by:

$$G^* = argmin_G max_D L_{cGAN}(G, D) + \lambda L_{L1}(G), \tag{3.16}$$

where

$$L_{cGAN}(G, D) = E_{x,y}[logD(x, y)] + E_{x,z}[log(1 - D(x, G(x, z))], \tag{3.17}$$

and

$$L_{L1}(G) = E_{x,y,z}[\|y - G(x, z)\|] \tag{3.18}$$

i.e., the generator is trained using the feedback from the conditional discriminator and Euclidean distance between the generated and original visible images.

Network Architecture

A defining feature of image-to-image translation problems is that they map a high resolution input grid to a high resolution output grid. In addition, for the paired images, the input and output differ in surface appearance, but both are renderings of the same underlying structure i.e., input structure is roughly aligned with the structure of the output.

Most of the generative models before pix2pix use encoder-decoder architecture (Johnson et al., 2016; Pathak et al., 2016; Ulyanov et al., 2016; Yoo et al., 2016; Zhou & Berg, 2016). In such a network, the input is passed through a series of layers that progressively downsample, until a bottleneck layer, at which point the process is reversed. Such a network requires that all information flow pass through all the layers, including the bottleneck. For many image translation problems, there is a great deal of low-level information shared between the input and output, and it would be desirable to shuttle this information directly across the net. For example, in the case of image colorization, the input and output share the location of prominent edges. To give the generator a means to circumvent the bottleneck for information like this, pix2pix adds

skip connections, following the general shape of a "U-Net" (Ronneberger et al., 2015). Specifically, they add skip connections between each layer i and layer $n - i$, where n is the total number of layers. Each skip connection simply concatenates all channels at layer i with those at layer $n - i$ as shown in Figure 3.9.

The discriminator is designed to only model high-frequency structure, relying on an L1 term to force low-frequency correctness (equation 3.18). In order ot model high-frequencies, it is sufficient to restrict out attention to structure in local image patches. Therefore, pix2pix discriminator is a *patchGAN*, that only penalizes structure at the scale of patches. This discriminator tries to classify if each $N \times N$ patch in a image is real or fake. This discriminator is convolutionally ran across the image, averaging all responses to provide the output of D. In this work, a 70×70 patch discriminator is used. This is advantageous because a smaller patchGAN has fewer parameters, runs faster, and can be applied to arbitrarily large images.

The discriminator is a Markovian discriminator or a PatchGAN - that only penalizes structure at the scale of patches. This discriminator tries to classify if each $N \times N$ patch in an image is real or fake. In this study a *70* $\times$ *70* patch discriminator is used. The other model used for this study is starGAN2, and this model is explained in detail in 3.1.

3.2.2 Photometric Normalization Techniques

Inspired from Whitelam and Bourlai, 2015, two basic photometric normalization techniques are applied to the original thermal and visible images before training and testing the models. These techniques are CLAHE (Contrast Limited Adaptive Histogram Equalization) and LBSSR (Log Based Single Scale Retinex). As a final technique, CLAHE is applied to the original images, followed by applying LBSSR.

CLAHE (Contrast Limited Adaptive Histogram Equalization)

CLAHE (Reza, 2004) operates on small local regions in the image and applies histogram equalization on each individual region (in contrast to the entire image in regular histogram equalization). In order to increase the amount of contrast while decreasing the amount of noise, CLAHE redistributes each histogram so that the height of each bin falls below a predetermined threshold. Specifically, gray levels below it. Finally, the patches are subsequently combined using bi-linear interpolation. Mathematically, this can be defined as

$$f(n) = \frac{N-1}{M} \times \sum_{k=0}^{n} h(k) \tag{3.19}$$

where M and N are the number of pixels and gray level bins in each sub-region, respectively, and h is the histogram of each sub-region.

SSR (Single Scale Retinex)

SSR photometric normalization technique (Jobson et al., 1997) decomposes the image into two components, illumination, *L(x,y)* (the amount of light falling on the target object) and reflectance, *R(x,y)* (the amount of light reflecting off the target object). The illumination is estimated as a low-pass version of the original image, while the reflectance component is obtained by dividing the original image from the illumination image. Mathematically, this can be described as:

$$I(x, y) = L(X, y) \times R(x, y) \tag{3.20}$$

$$L(x, y) = I(x, y) * G_{\sigma}(x, y) \tag{3.21}$$

where G_σ is a Gaussian of scale σ and * denotes the convolution between the image and the kernel. Finally the reflectance of the image id estimated as:

$$R(x, y) = \log \frac{I(x, y)}{L(x, y)} \tag{3.22}$$

3.3 Involution GAN: Rethinking Architecture to Improve the Performance of Cross-Spectral Face Recognition

The methodological approach followed to solve the problem of cross-spectral face recognition is as follows: First, ArcFace model is trained to establish a baseline. Then, pretrained ArcFace is used to extract 512-D feature embeddings from the original visible and thermal face images, which are saved to be used later. Then, StarGAN2 model is trained in its original form to establish a synthesis baseline, and then the proposed InGAN (Involution-GAN) is trained. For the next set of experiments, an identity loss term is incorporated in the form of cosine distance between the embeddings generated from the original visible images before training and the embeddings generated from the synthesized thermal and visible images (since the GAN is cyclic). Then, the trained models are used to generate visible band images from the thermal acquisitions in the test set. ArcFace is them used to match the original and the synthesized visible face images. The methodology is illustrated in detail in Figure 3.10

3.3.1 Involution

The concept of involution is introduced in D. Li et al., 2021, which is described as inverting the inherence of convolution (not to be confused with deconvolution, which is the true inverse operation of convolution). Convolution is, by and large, an operation that is spatial-agnostic and channel-specific. The former property can be explained as using the same filter for convolution across spatial dimensions. Channel-specific indicates different filters being used for each of the channels. These properties carry the following disadvantages: filters won't be able to adapt to the diverse visual patterns with respect to different spatial

positions; and a lot of redundancy across the channels. Convolution operation in an image context is shown in Figure 3.11.

To address these problems, involution is spatial-specific to be able to learn diverse visual patterns across spatial positions, and it is channel-agnostic, i.e., the same filter is used at a given position throughout the depth of the input. To further explain the functionality, convolution operates by sliding the same kernel over the entire 2D image, and a different kernel is slid over the next layer. In involution, filters are dynamically generated at every spatial position based on the neighborhood. These spatial-specific filters are then broadcasted across all the channels. This reduces the training time by reducing the number of training parameters and since these parameters depend on the spatial location, training converges quicker. Involution operation is illustrated in Figure 3.12.

3.3.2 *Redblk*: Residual Network using Involution

Deep CNNs have a led to a series of breakthroughs for image classification. Now, as these deeper networks start converging, a degradation problem has been exposed. This degradation implies that increasing the depth is not always the solution. To address this, *deep residual framework* (K. He et al., 2016) is introduced where instead of trying to fit each few stacked layers to fit a desired underlying mapping, the layers are explicitly fit to a residual mapping. Mathematically, consider $H(x)$ as an underlying mapping that needs to be fit by a few stacked layers, where x is the input to the first of these layers. The stacked layers are enabled to fit another mapping $F(x) := H(x) - x$. Now, the original mapping becomes $H(x) = F(x)+x$. This formulation of $F(x)+x$ can be realized by feed-forward networks with "shortcut connections", which means to skip one or more layers. In this case, these skip connections simply perform identity mapping, and their output is added to the output of the stacked layers, as shown in Figure 3.13. These identity shortcut connections add neither to the number of parameters, nor to the computational complexity. The entire network can still be trained end-to-end using stochastic gradient descent (SGD) with backpropagation, and can be easily implemented using common deep learning libraries without the need to modify any solvers.

3.3.3 Proposed InGAN Architecture

Neural network blocks using involution as the atomic operation are built and skip connections or residuals (K. He et al., 2016) are used, and these are called *RedBlks*, following RedNets in D. Li et al., 2021. The RedBlk built using the involution operation is visualized in Figure 3.13. The three blocks in this InGAN model are constructed using these RedBlks. The detailed network architecture of each of the building blocks is shown in Tables 3.1 and 3.2 and in Figures 3.14 and 3.15. All the downsampling blocks and the first two blocks of the bottleneck in the generator use instance normalization (Ulyanov et al., 2016), while the last two bottleneck blocks and all the upsampling blocks use Adaptive Instance Normalization (AdaIn) (X. Huang & Belongie, 2017; Karras et al., 2019).

Table 3.1: Generator architecture, includes downsampling, upsampling, and bottleneck blocks.

Layer	Output shape	# blocks	Functionality
Input image	256×256×3	-	-
Inv1×1	256×256×64	1	Downsample
Inv3×3 ReLU Inv3×3 ReLU AvgPool	16×16×512	4	Downsample
Inv3×3 ReLU Inv3×3 ReLU	16×16×512	4	Bottleneck
Inv3×3 ReLU Inv3×3 ReLU	256×256×64	4	Upsample
Inv1×1	256×256×3	1	Upsample

Table 3.2: Discriminator and style encoder architecture. In the dense layer, K=2 for the discriminator for the two possible outcomes (real/fake). K=64 for the style encoder, for the output dimension of the style code. Both, the style encoder and the discriminator have two output branches, one for each of the domains

Layer	Output shape	# blocks	Functionality
Input image	256×256×3	-	-
Inv1×1	256×256×64	1	-
Inv3×3 ReLU Inv3×3 ReLU AvgPool	4×4×512	6	Downsample
LReLU	4×4×512	1	-
Inv4×4	1×1×512	1	Downsample
LReLU	1×1×512	1	-
Flatten	512	1	Reshape
Dense	D × 2	1	output

3.3.4 Training Objective

The GAN training objective is given below.

$$\min_{\text{G,F,E}} \max_{\text{D}} L_{adv} + \lambda_{sty} L_{sty} + \lambda_{cyc} L_{cyc} \tag{3.23}$$

where L_{adv} is the adversarial loss, L_{sty} is the style reconstruction loss, and L_{cyc} is the cyclic loss (Zhu et al., 2017), and λ_{sty}, and λ_{cyc} are the hyper-parameters for each term. For more details on the objective and loss terms, please refer (Choi et al., 2020).

Reducing the inter-identity distance is proved to improve the face verification accuracy when using the synthesized images (Zhang et al., 2017) (Chen & Ross, 2019). To achieve this, pre-trained ArcFace model is used to extract features in the form of 512-D embeddings from the original face images prior to training the GAN model. During training, ArcFace is again used to generate embeddings from the synthesized images after each iteration. Then, the cosine distance between these embeddings and those saved prior to training (that belong to the same identity) is calculated and is used as identity loss and is shown below.

$$\mathcal{L}_{\text{id}} = cos\left[Emb_{G(x)}, Emb_y\right] \tag{3.24}$$

where $G(x)$ is the generated image from the source image, x, and y is the corresponding reference image, and Emb indicates the 512-D embedding obtained using ArcFace.

The new objective to train InGAN is now,

$$\min_{\text{G,F,E}} \max_{\text{D}} L_{adv} + \lambda_{sty} L_{sty} + \lambda_{cyc} L_{cyc} + \lambda_{id} L_{id} \tag{3.25}$$

where λ_{id} is the hyperparameter for the identity loss.

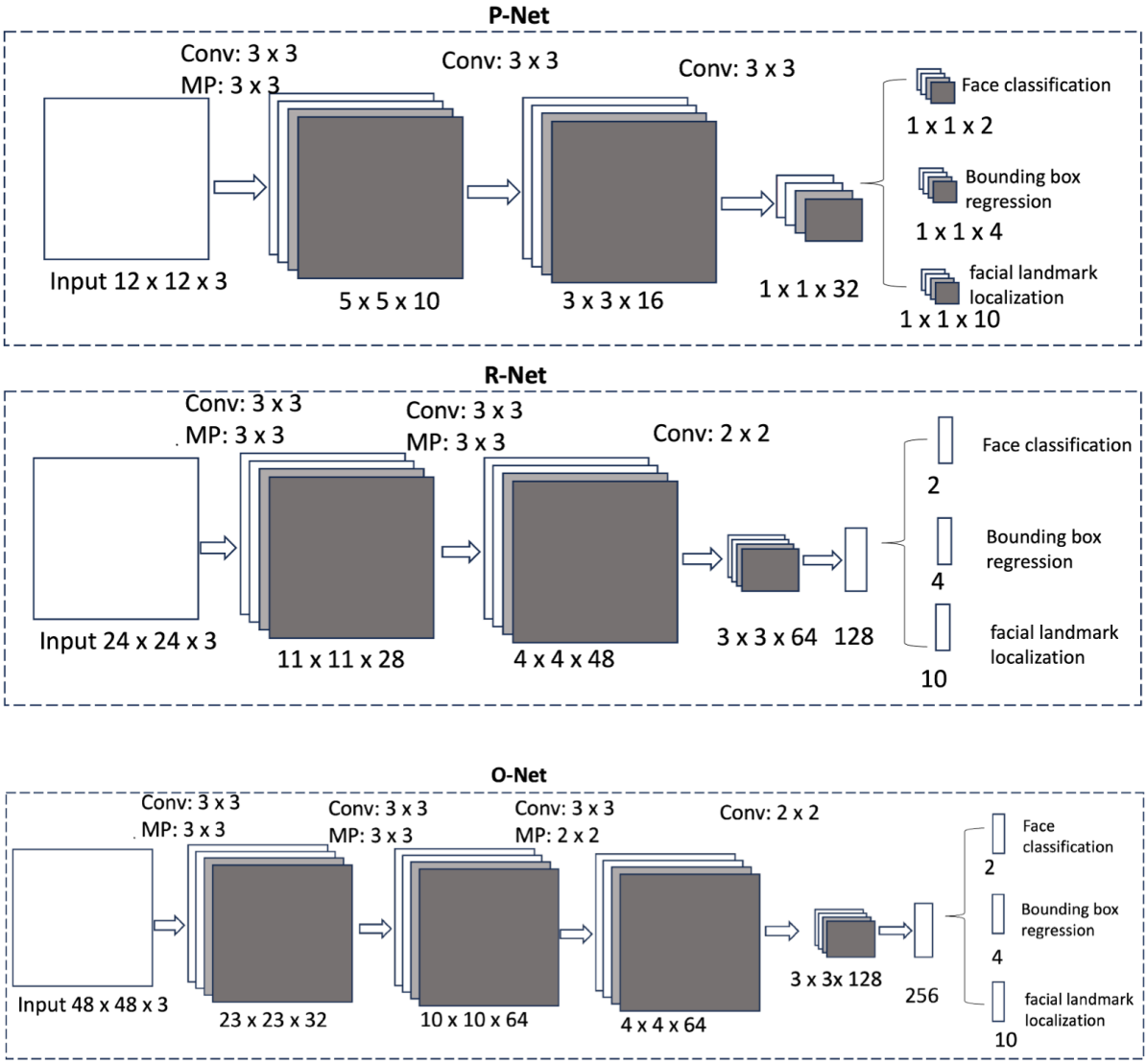

Figure 3.2: Calculating normalized error from the predicted and actual eye centers, adopted from Jesorsky et al., 2001. P_R and P_L are the predicted right and left eye centers respectively, and A_R and A_L are the actual right and left eye centers respectively. Normalized error is the the worst eye estimation divided by the Euclidean distance between actual eye centers.

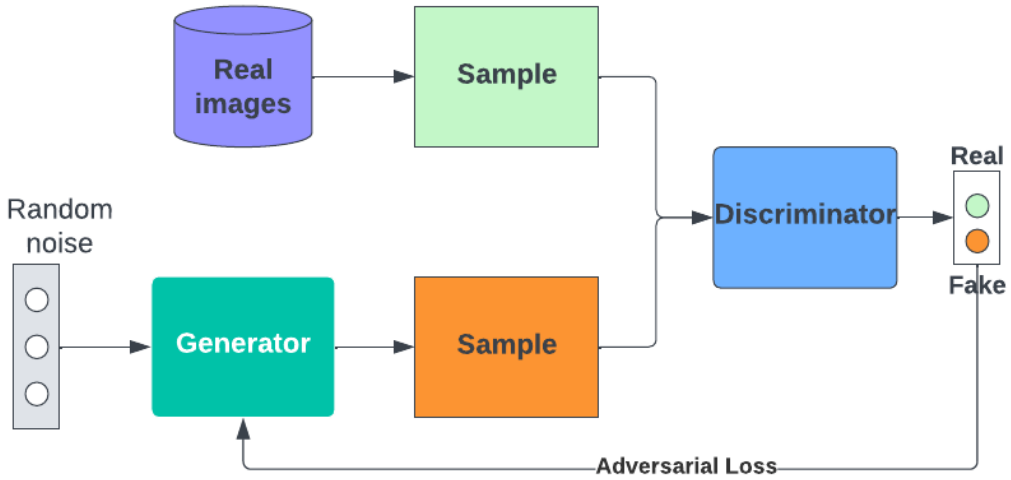

Figure 3.3: Figure showing the functionality of a vanilla GAN. Generator accepts random noise as input and tries to generate images that look similar to the real images. The discriminator takes in the real and fake images as inputs and tries to distinguish. A GAN can be viewed to have trained successfully when the discriminator cannot distinguish between the real and fake images.

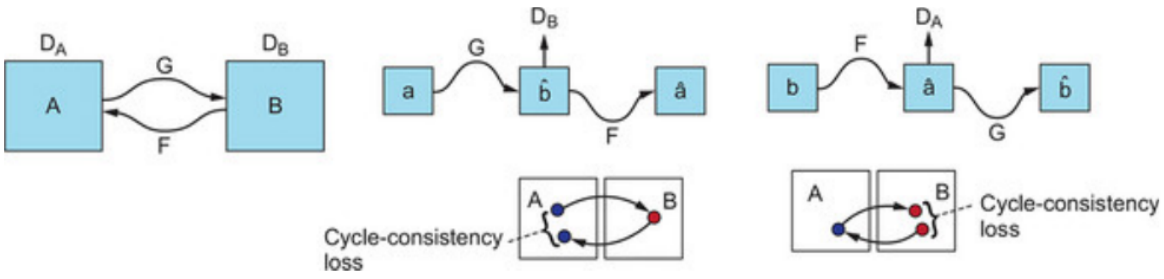

Figure 3.4: CycleGAN consists of two mapping functions $G : X \rightarrow Y$ and $F : Y \rightarrow X$, and associated adversarial discriminators, D_X and D_Y. It is trained by optimizing two adversarial losses and two cycle consistency losses together.

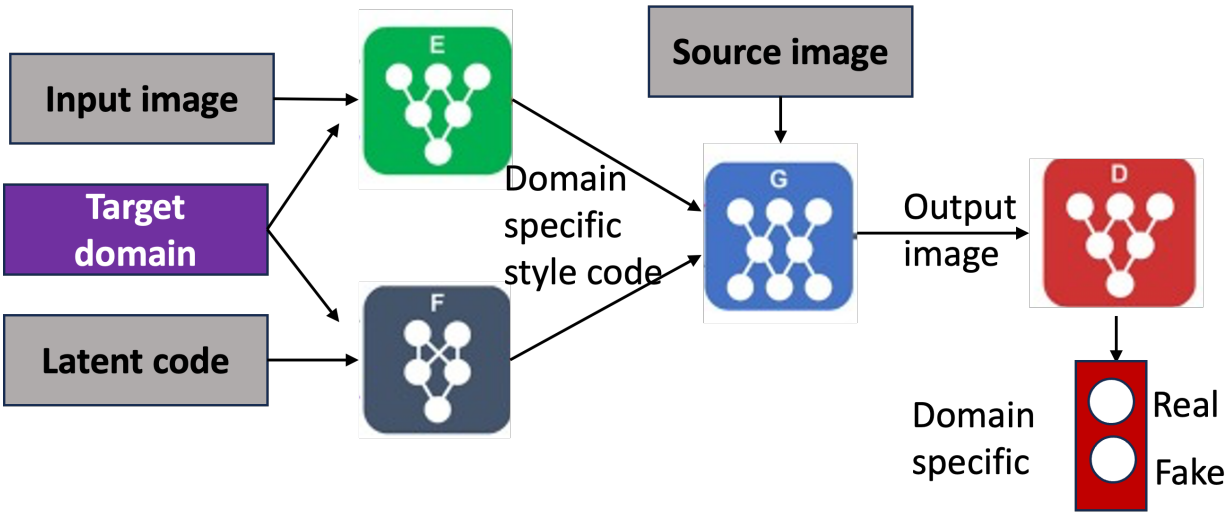

Figure 3.5: StarGAN consists of four different networks: A style encoder that generates a style code to match a reference image, used for reference-guided synthesis; A mapping network that generates style code from a domain, used for latent-guided synthesis; A generator that takes style code, either from the mapping network or the style encoder to generate images; and lastly, a discriminator network to distinguish between the real and the generated images.

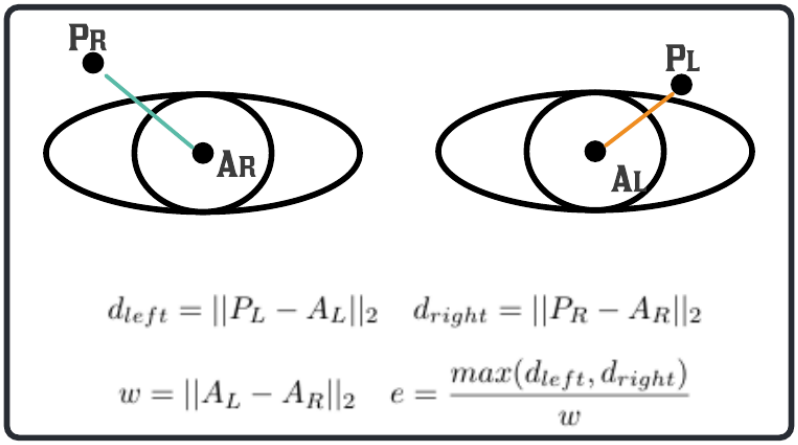

Figure 3.6: Calculating normalized error from the predicted and actual eye centers, adopted from Jesorsky et al., 2001. P_R and P_L are the predicted right and left eye centers respectively, and A_R and A_L are the actual right and left eye centers respectively. Normalized error is the the worst eye estimation divided by the Euclidean distance between actual eye centers.

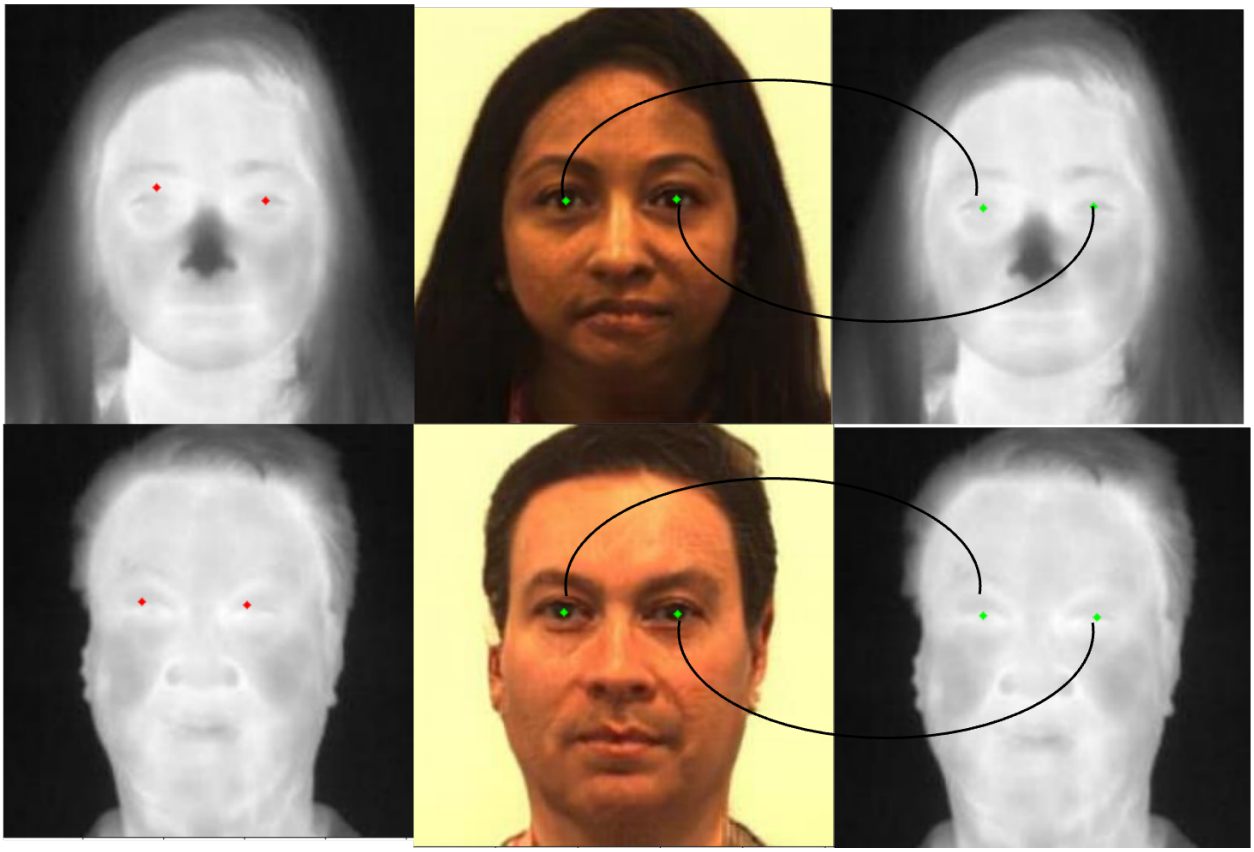

Figure 3.7: MT-CNN is used to detect the eye centers in the synthesized visible images, then these coordinates are mapped to the corresponding original thermal images (green labels). The red labels are from the baseline model.

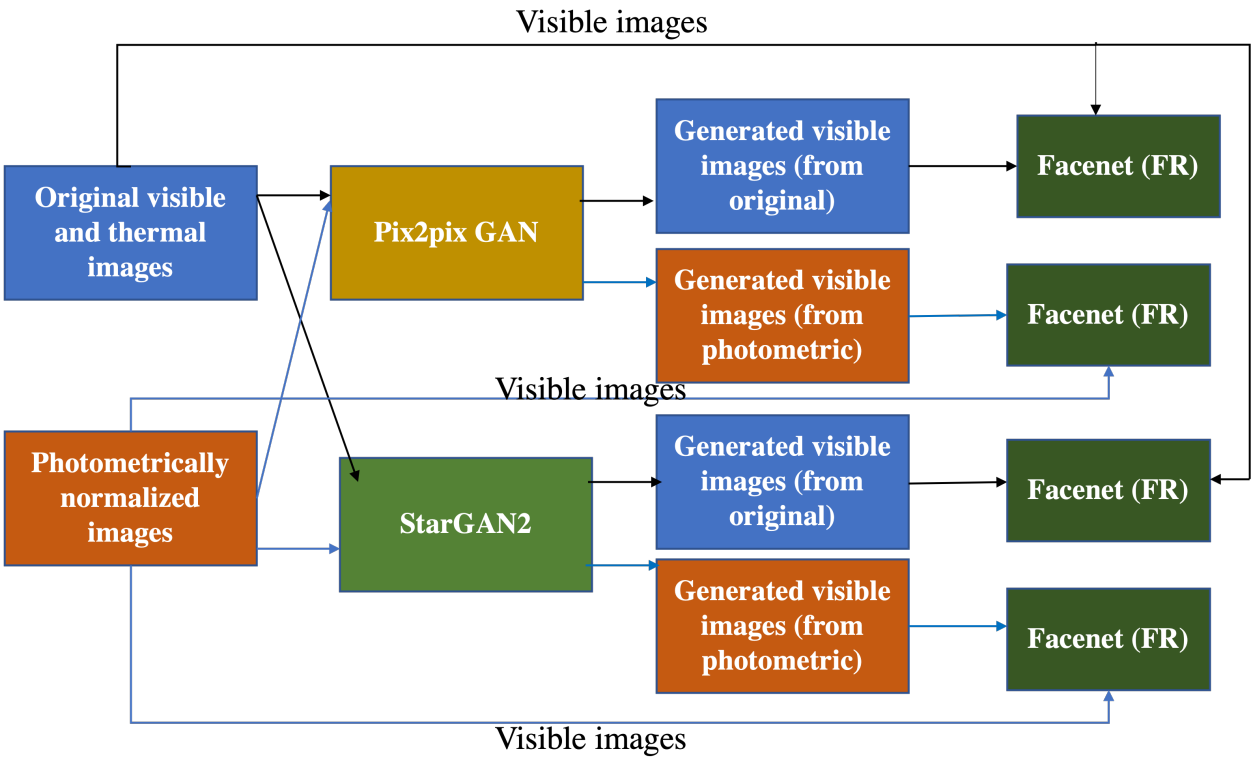

Figure 3.8: Methodological approach followed in this paper. Two variations of GANs (Pix2pix and StarGAN2) are trained using first the original thermal and visible images, and then using photometrically normalized images. In the test phase, after the images are generated using the trained models, original and synthesized visible images are used for face recognition using Facenet (pretrained visible-to-visible FR model).

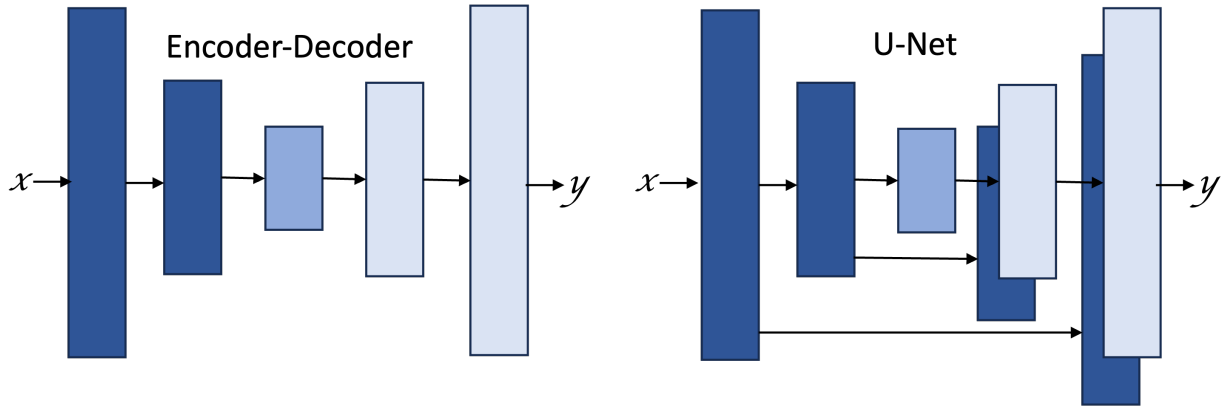

Figure 3.9: Figure explaining the difference between the regular encoder-decoder architecture and the U-Net. The skip connections in the U-Net provide a means to the generator to circumvent the bottleneck for a great deal of low-level information.

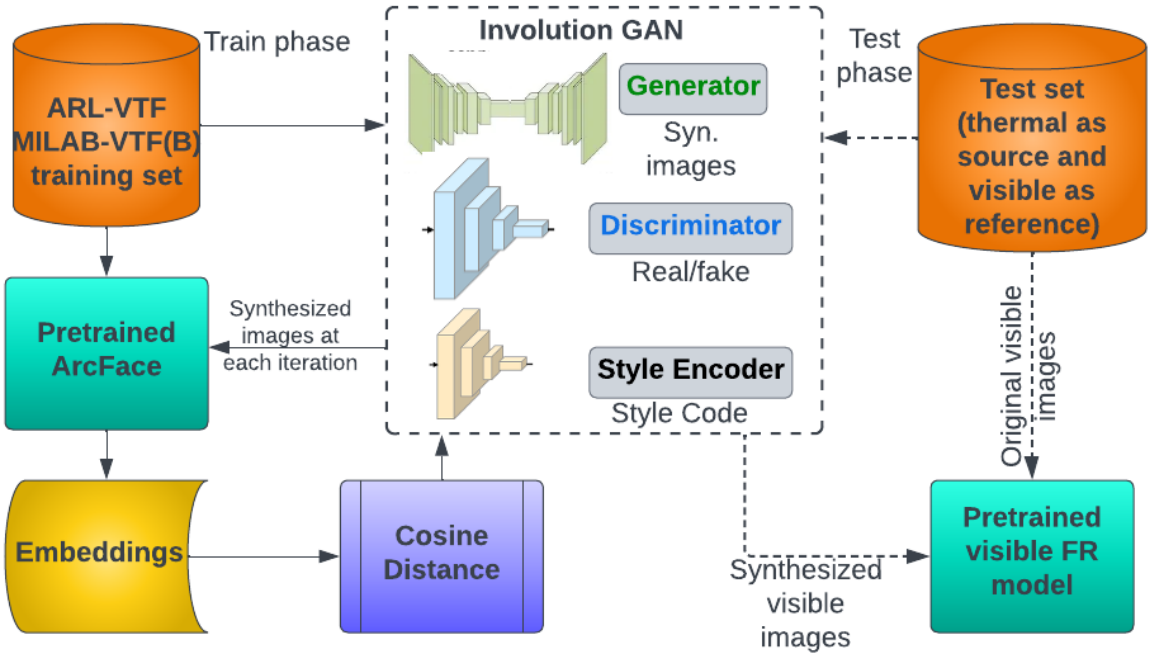

Figure 3.10: The methodological approach followed in this work is shown here. Succinctly, during training (represented by the solid line flow), ArcFace is used to extract embeddings from the original and the synthesized images at every iteration to compute the identity loss, Involution GAN is trained. Then the trained model is used to synthesize visible images from the thermal images during the test phase (represented by the dashed line flow). Next, these synthesized and original visible images are matched using a pretrained visible FR model.

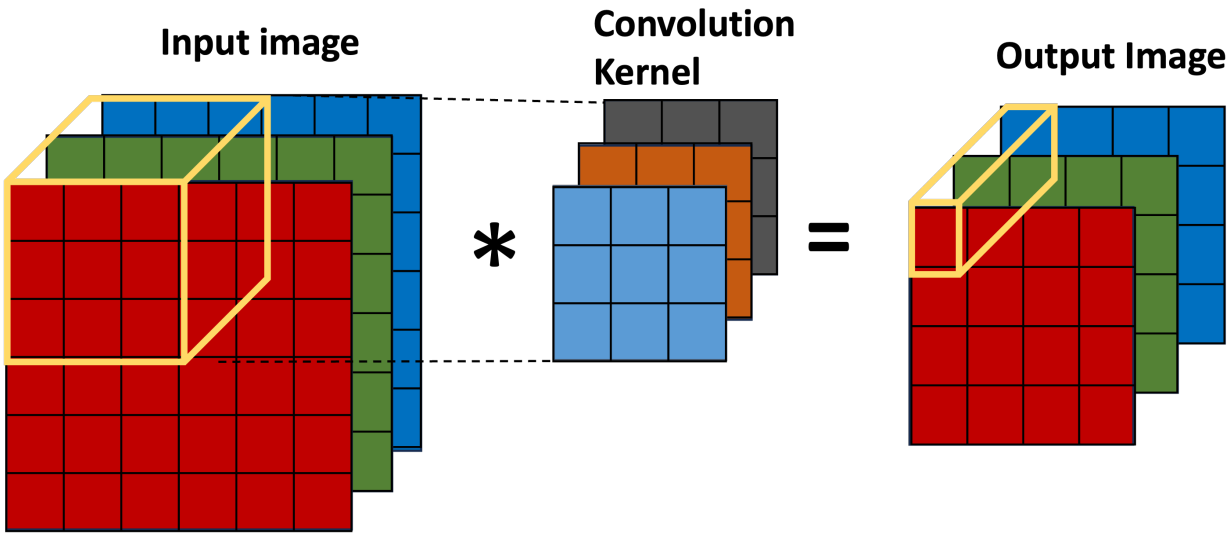

Figure 3.11: Convolution is a fairly simple operation. Each 2D kernel (for instance, blue, in the figure) slides over the input data, performing an element-wise multiplication with the part of the image the input is currently on (yellow cube on the input image), and then adding the results into a single output pixel (yellow cube on the output image).

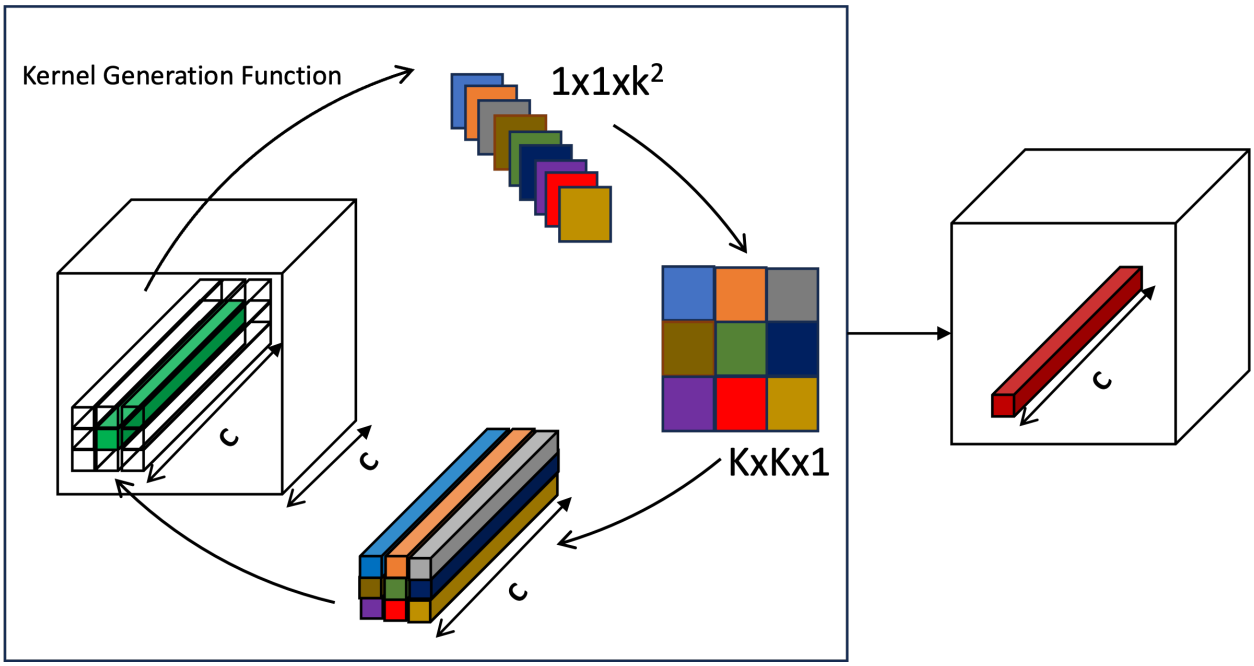

Figure 3.12: Involution operation and RedBlk built using it are shown here. The involution kernel ($1 \times 1 \times k^2$) is yielded using a kernel function conditioned on a single pixel at that spatial location, followed by a channel-to-space rearrangement converting it to the shape $k \times k \times 1$. Then, this kernel is broadcasted across all the channels for that pixel neighborhood.

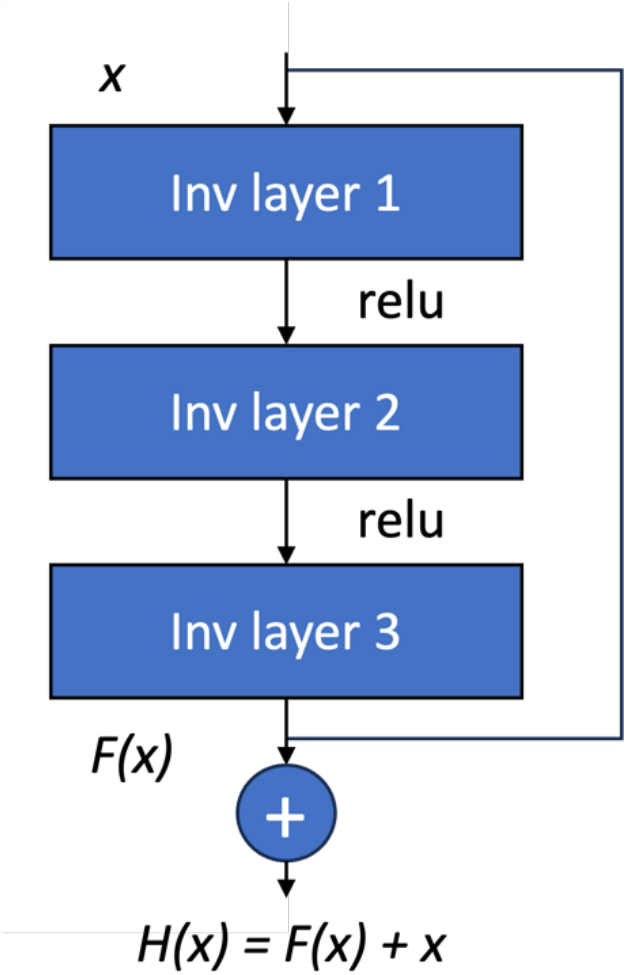

Figure 3.13: The *RedBlk*

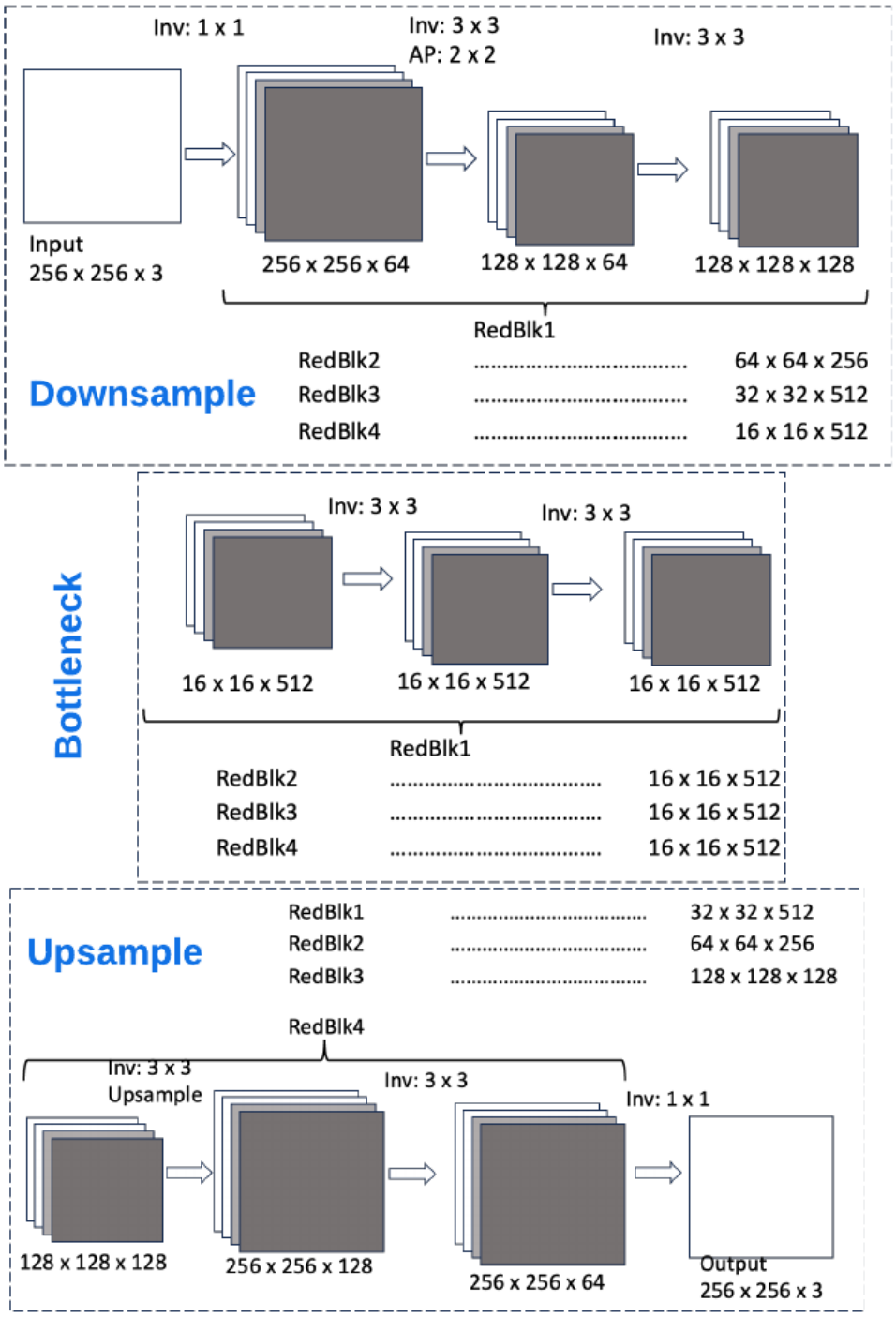

Figure 3.14: Figure showing the architecture of the proposed InGAN generator. It can be considered as an ensemble of three structures. The first is the upsample network which includes four RedBlks, second, a bottleneck with four blocks, and third is a downsample network with four blocks.

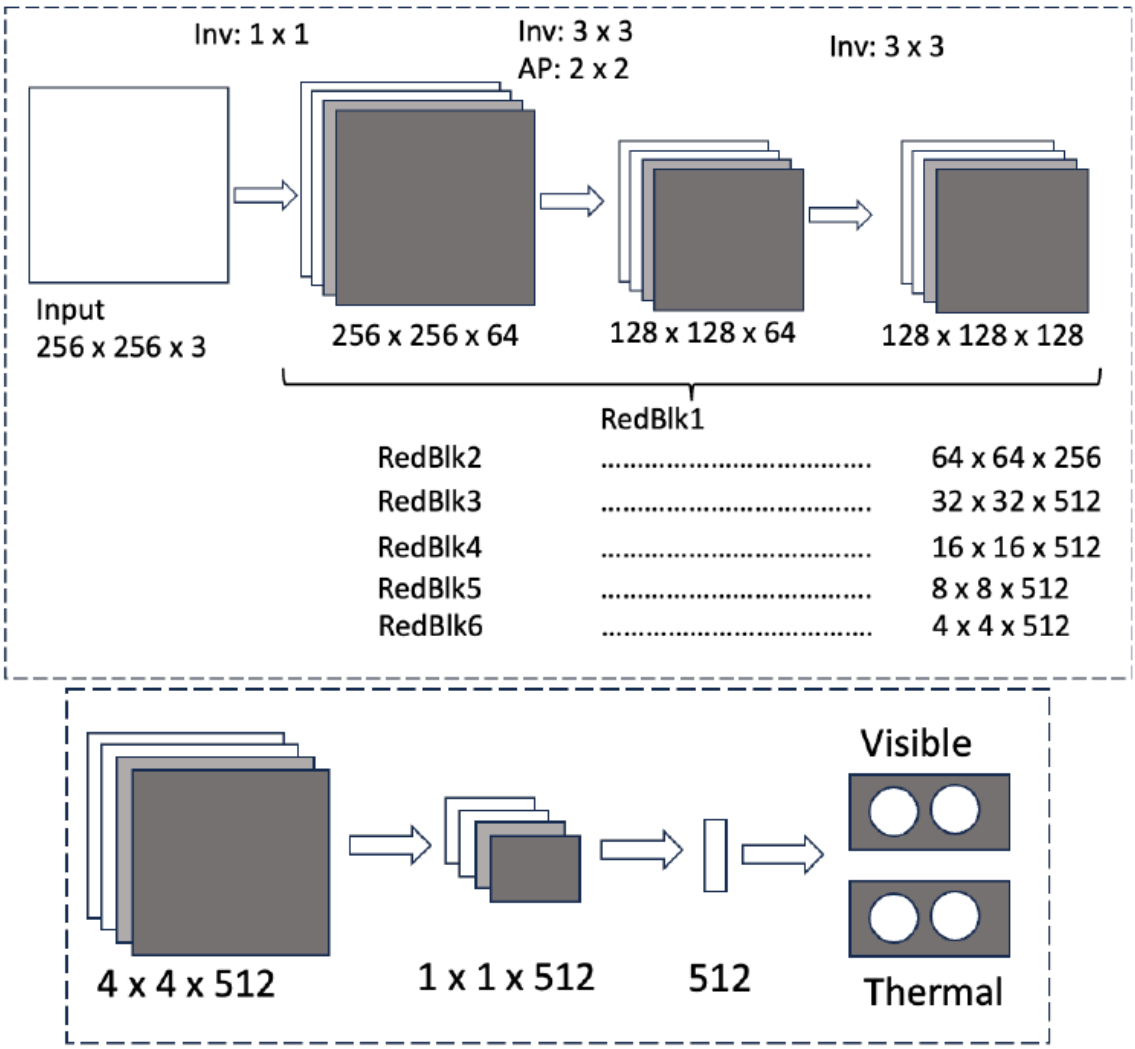

Figure 3.15: Figure showing the architecture of the proposed InGAN discriminator. It includes 6 redblks followed by an involution layer and a fully connected network, resulting in the number of output branches equal to the number of domains. Each of the output branch has two nodes, one for real and one for fake output.

Chapter 4

Experimental Evaluation and Results

This chapter explains the datasets used, experimental protocol followed, and the results obtained in great detail for each of the proposed solutions

4.1 Datasets used

Two datasets are used to perform the experiments in this study: Army Research Laboratory - Visible Thermal Face (ARL-VTF) dataset (Poster et al., 2021) and Multi-spectral Imagery Lab - Visible Thermal Face (MILAB-VTF(B)) dataset (Bourlai et al., 2023).

4.1.1 ARl-VTF dataset

This dataset consists of mid-wave infrared (MWIR) and visible images from 395 subjects, out of which the data from 295 subjects is used for training and the remaining 100 for testing. It includes full frontal images with neutral expression, pose images, and images with expression. The baseline sequence is collected with the subjects looking directly at the camera with a neutral expression. The pose sequence of images is collected by asking the subjects to slowly turn their heads from left to right. The expression sequence of frontal images is collected by asking the subjects to count out loud incrementally starting from one. It also includes an additional baseline where subjects were asked to wear glasses if they typically do.

For the eye center detection problem, 10 images for each subject from each band are used for training. This resulted in a total of 5900 training images. For testing, 5 frontal images from each subject are used per band, resulting in a total of 500 test images per band. Images in this band do not have much variation in terms of in-plane rotation. Therefore, to understand the effects of in-plane rotation on synthesis, and thereby on eye center detection, all the training and test images are rotated clock-wise and anti clock-wise at angles 15 and 30 degrees. This increased the size of the training and test set by five fold.

StarGAN2 requires images of size 256×256. Therefore, all the training and test images are cropped using the given ground truth bounding boxes to a scale of 1.7 to include a consistent margin around the face, as shown in Figure 4.9. CycleGAN is trained and tested using the same dataset as described above.

For the cross-spectral face recognition experiments, the same training set as described is used. For testing, 1 frontal image for each subject per band is used, resulting in a total of 200 test images. Then, the more challenging MILAB-VTF(B) dataset described below is used to run the experiments.

4.1.2 MILAB-VTF(B) Dataset

This section describes the process followed when the MILAB team at UGA collected this MWIR-Visible dataset.

Data Collection

Informed Consent and IRB Procedure. To be able to de-identify the data and avoid duplicate data, the participants were assigned a random identification number associated with their biometric data after completing the registration. The participants were also asked to consent or decline the use of their biometric images in publications, i.e., their facial images can be used as examples in future publications that may include but are not limited to, research papers, journal articles, presentations, educational material, or other related documents. The study involved an approximately 40-minute 2 session process, on the same day. First, an indoor session is carried out in a black camping tent that was reinforced internally so that limited light was coming in. Next, an outdoor session, that involved facial image and video captured at short and long stand-off distances is performed. The state-of-the-art sensors used are from Canon (Mark IV), Nikon (PX1000) and two FILR sensors, one of which is short range, and the other, long range (see Table 4.1).

Indoor Session. Every subject first completed an Institutional Review Board (IRB) consent form before beginning the collection process. After completing the IRB form, the participants were taken inside the tent one at a time. Three images were captured using the Canon and A8581 cameras; (1) full frontal with the subject facing the cameras, (2) full left profile, and (3) full right profile. Next, a video was captured where the participants were instructed to turn their head, starting from a full frontal position to a full left profile, then to a full right profile, then look up and then down before returning to a full frontal pose at the end of the video.

Outdoor Session. After completion of the indoor session, the participants were taken outside to the collection area outside the tent. Participants were instructed to walk to each of the outdoor collection areas at 100, 200, 300, and 400 meters from the camera. Videos of each subject were recorded with the Nikon P1000 and FLIR 8512 cameras at each distance. Participants were instructed to turn their head as they did in the indoor setting in each of the four different distances. Data was collected in a variety of weather conditions.

Table 4.1: Camera Equipment. State-of-the-art camera sensors are used to capture high-resolution MWIR and visible-spectrum images of subjects at various standoff distances. The exact camera configurations below are specified below. The Cannon Mark IV, Nikon P900, and FLIR A8582 are used for indoor data collection, while the Nikon P1000 and FLIR RS8513 are used for outdoor data collection.

Camera	Spectrum	Spectral Range	Focal Length	F-Stop	Resolution
Canon Mark IV	Visible	-	70-200mm	2.8-32	1920×1080
Nikon P900	Visible	-	24-2000mm	2.8-6.5	1920×1080
Nikon P1000	Visible	-	24-3000mm	2.8-8	3840×2160
FLIR A8581	MWIR	3.0 - 5.0 μm	50m	2.5	1280×1024
FLIR RS8513	MWIR	3.0 - 5.0 μm	120-1200mm	5	1280×1024

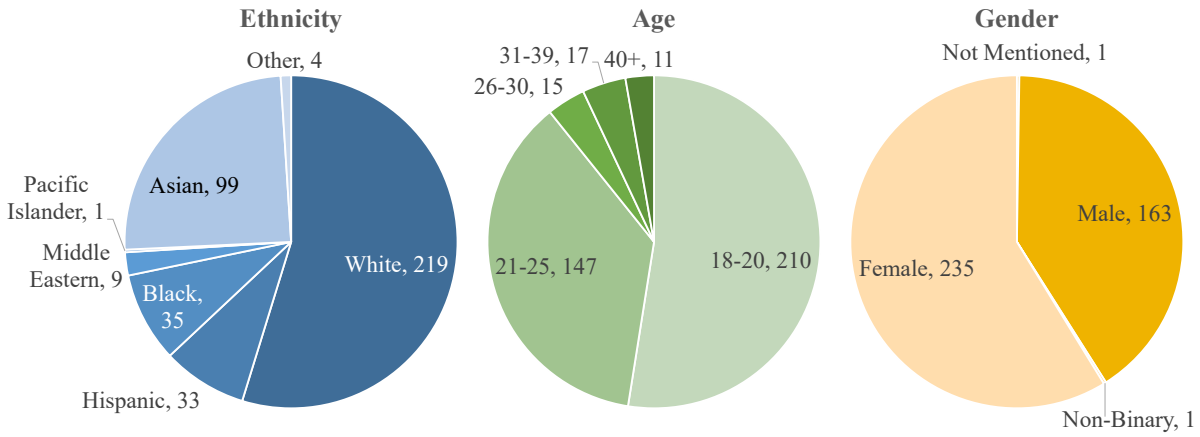

Figure 4.1: The MILAB-VTF(B) dataset is diverse with respect to ethnicity, age, and gender.

Data Structure - Training and Evaluation Protocol

The MILAB-VTF(B) dataset provides unsynchronized, paired thermal-visible videos and anonymized identifiers for each subject. Algorithmically generated frame synchronization between thermal and visible videos, face bounding boxes, and facial key points, which will be useful for developing end-to-end multi-spectral face verification pipelines are provided. Out of the 400 subjects, 320 subjects are selected for training, and sequester 80 identities are used for evaluation. Following standard face verification protocols, gallery and query sets are created from the sequestered data. Specifically, four non-overlapping galleries and four non-overlapping query sets are created by splitting the evaluation data by pose (i.e., frontal/profile) and location (i.e., indoor/outdoor).

In addition to the algorithmically generated face bounding boxes and key points, A small subset of the dataset is labeled manually to allow for the additional evaluation. Five images from each distance,

location, and spectrum for all 400 subjects are selected and labeled based on specific poses (frontal, left profile, right profile, facing up and facing down). For each image, a face bounding box and seven landmarks are annotated. The landmarks include the inside and outside corners of both eyes, tip of the nose, and the left and right mouth corners.

4.1.3 Benchmark Experiments using MILAB-VTF(B)

This section provides details on the experiments performed to establish baseline results using the MILAB-VTF(B) dataset. The experiments performed are multi-spectral face detection, and visible-to-visible same distance face recognition.

Multi-spectral Face Detection. Google's Tensorflow object detection API (J. Huang et al., 2017) is used to perform the face detection experiments. To perform these experiments, a subset of the dataset, which is manually annotated for face bounding boxes is used. The train set in the manually annotated dataset consists of five images per subject, per distance, and per spectrum. the fie image are of the subject looking towards the camera, full left profile, full right profile, looking up, and looking down. the test set follows the same structure with images from 80 subjects, resulting in a total of 4000 images.

The object detection model used for this study is the Faster R-CNN with Resnet-50 as the feature extractor (J. Huang et al., 2017). This open-source model can be downloaded along with the pre-trained weights on the COCO dataset (Lin et al., 2014). This model is first fine-tuned on the WIDER face dataset (Yang et al., 2016) with a learning rate of 0.00001 and is trained to a total of 150,000 steps with a batch size of 4. The resultant model is then fine-tuned on the manually annotated training set with the learning rate of 0.0003 for a total of 40,000 steps with a batch size of 1. The precision and recall metrics obtained when this model is evaluated on the test set are presented in Tables 4.2 and 4.3, and the resultant face bounding boxes on the test images are shown in Figure 4.3.

Table 4.2: Face Detection Results. The test includes images from thermal and visible bands at all the distances. AP is the average precision over IOU's 0.5:0.95. AP_s, AP_m, AP_l are the average precision values over small, medium and large objects respectively.

Distance	AP	AP@.50	AP@.75	AP_s	AP_m	AP_l
Indoor	77.6	99.0	95.5	-	-	77.6
100 Meter	81.8	99.0	96.1	-	-	81.8
200 Meter	80.9	100	98.9	-	-	80.9
300 Meter	77.7	100	96.9	-	-	77.7
400 Meter	48.6	85.4	49.4	16.3	19.8	74.3
All	66.7	95.7	76.2	15.9	41.5	77.7

Same Spectral Face Recognition This section explains the same-spectral face recognition experiments performed using the MILAB-VTF(B) dataset. Face recognition is performed on visible images, where the gallery and query images are frontal and are collected at the same distance. Next, MTCNN (Xiang & Zhu, 2017) is used to detect eye centers in all the images. These eye centers are then used to geometrically normalize the images. These normalized images are shown in Figure 4.4.

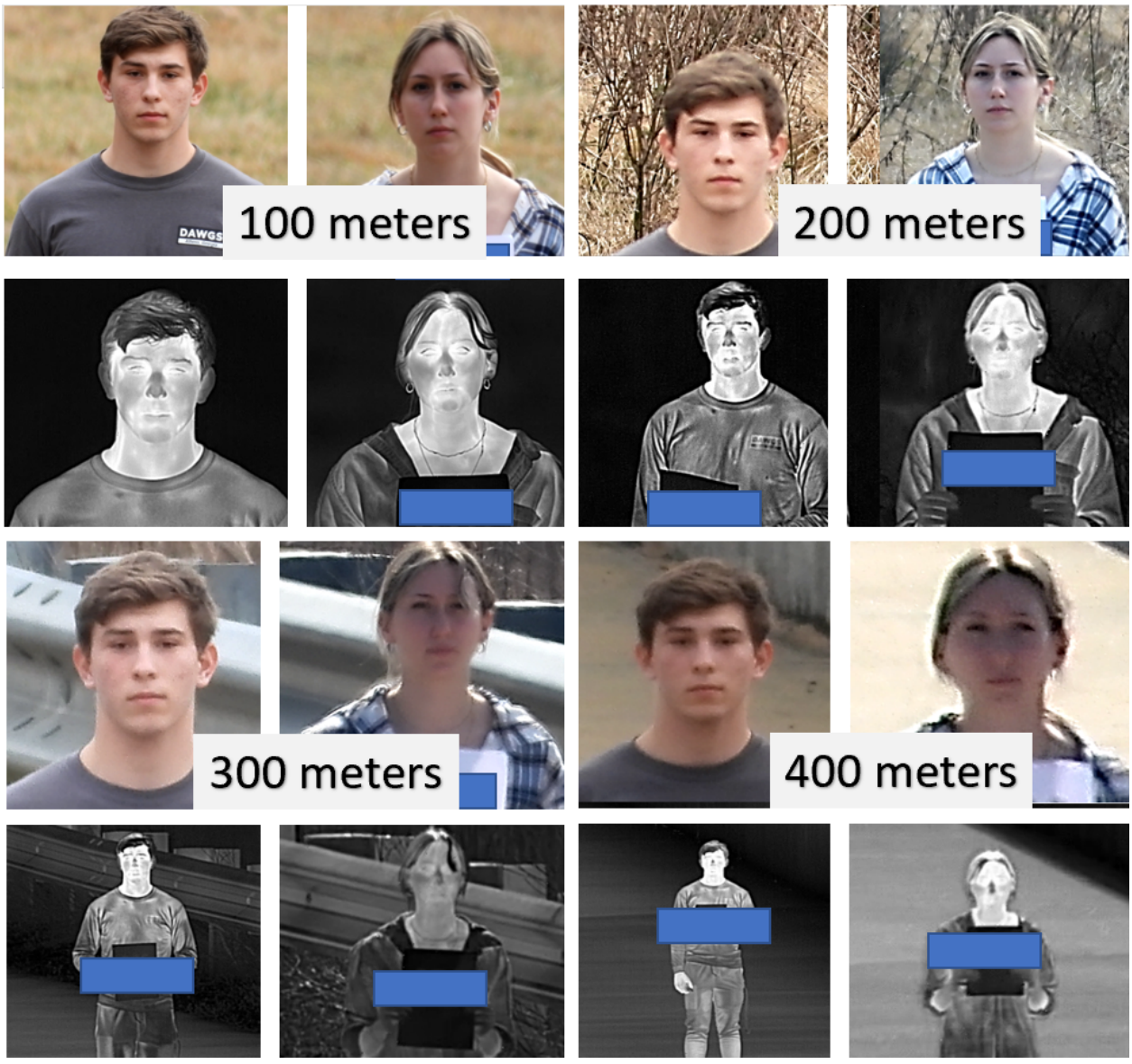

Figure 4.2: Outdoor sample face images from two different subjects pertaining to the MILAB-VTF(B) dataset collected during the COVID-19 era. The original (raw) face images shown are collected at four different distances, 100, 200, 300, and 400 meters.

Table 4.3: Face Detection Results. The test includes images from thermal and visible bands at all the distances. AR represents average recall and follows the same nomenclature as the precision metrics.

Distance	AR	AR_s	AR_m	AR_l
Indoor	83.2	-	-	83.2
100 Meter	85.0	-	-	85.0
200 Meter	84.1	-	-	84.1
300 Meter	81.0	-	-	81.0
400 Meter	53.3	24.0	34.3	78.2
All	70.5	23.7	50.2	82.2

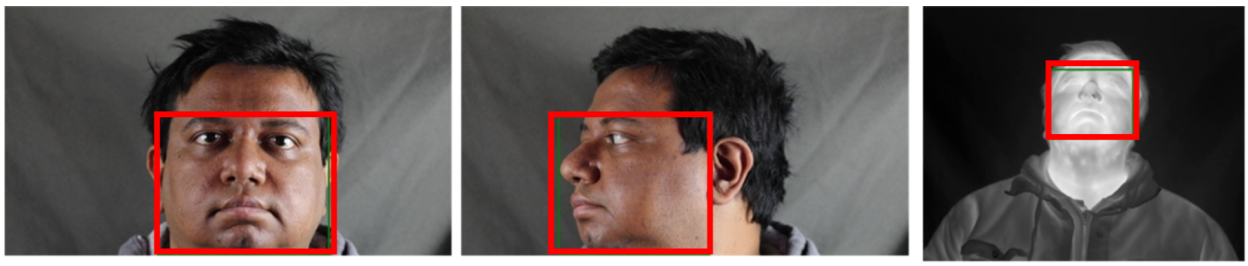

Figure 4.3: Indoor images with the detected face bounding boxes. The detector performed well for indoor images.

Three popular and efficient pre-trained FR models are used for this work, namely, Facenet (Schroff et al., 2015), ArcFace (Deng et al., 2019), and VGG-Face(Parkhi et al., 2015). Face verification metrics presented are Area under curve (AUC), Equal error rate (EER), True acceptance rate (TAR) @ False acceptance rates (FAR) 1% and 5%. As per the protocol, 80 subjects are used for testing. Since these experiments are performed using pre-trained models, training data is not needed.

The performance of all the models decline as the distance increases. An exception is the performance when going from 300 to 400 meters. The performance at 400 meters is consistently and slightly higher than the performance at 300 meters for all the models. This could be explained by the zoom-in coefficient being higher at 400-meter images during data collection. Also, the performance of ArcFace declines significantly at 200, 300, and 400 meters compared to the other models. This may be because of the small window size of the model, which is 112×112 pixels. VGG-Face performed the best at all the distances except for indoor, and the window size for this model is 224×224 pixels. Face verification metrics using VGG-Face are presented in Table 4.4, and the ROC curves using all the three models are shown in Figures 4.5, 4.6, and 4.7.

Figure 4.4: Geometrically normalized images. Top row - indoor, 100 and 200 meters. Bottom row - 300 and 400 meters.

4.2 Utilizing Alignment Loss to Advance Eye Center Detection and Face Recognition in the LWIR Band

This section explains the experimental protocol followed to address the eye center detection problem in the thermal spectrum.

4.2.1 Baseline Experiments

MT-CNN (Xiang & Zhu, 2017) is trained using thermal images and corresponding bounding boxes and landmarks to establish a baseline. MT-CNN needs four different kinds of annotations for the training dataset.

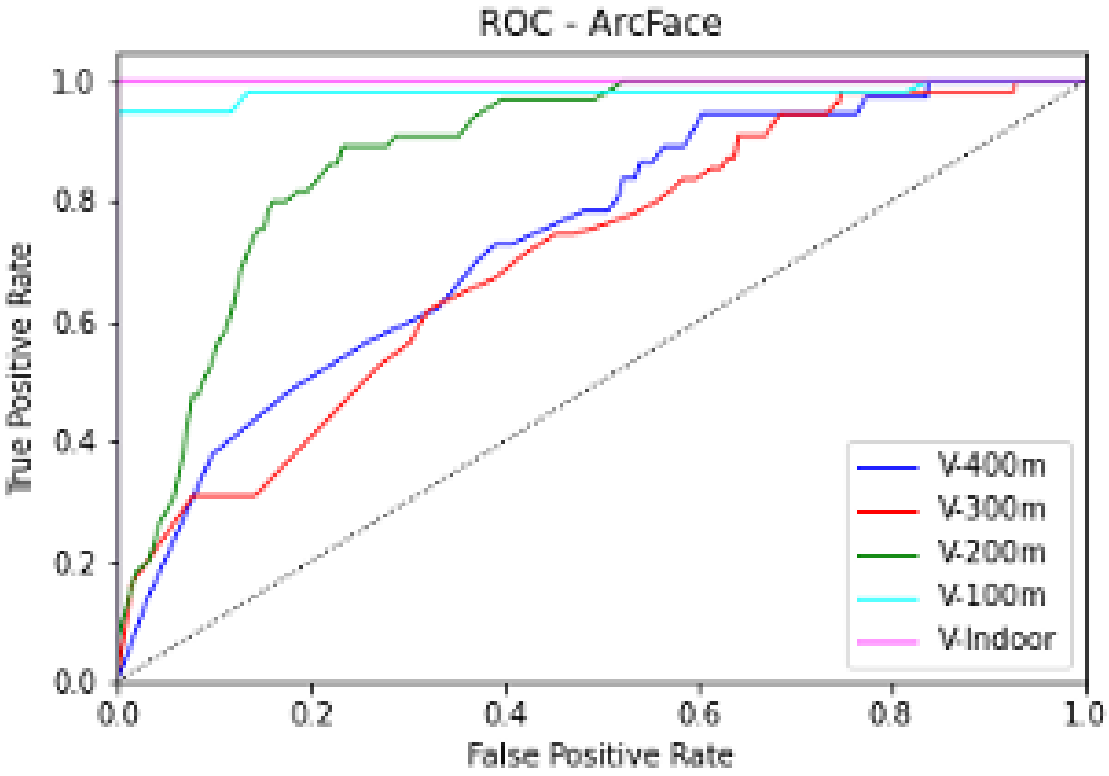

Figure 4.5: ROC curves for same spectral face recognition using ArcFace.

- Negatives: Regions where the Intersection-over-Union (IoU) ratio is less than 0.3 to the ground-truth face,
- Positives: IoU above 0.65 to a ground-truth face,
- Part faces: IoU between 0.4 to 0.65 to a ground-truth face,
- Landmark faces: Faces labeled with 5 landmark positions

The authors of MT-CNN use WIDER FACE (Yang et al., 2016) to collect positives, negatives, and part face, and CelebA (Z. Liu et al., 2018) as landmark faces. For this study, the same dataset described in the dataset section for all the annotations is used. Several patches from each image are cropped to collect positives, negatives, and part faces to train the P-Net. The detected faces from the first stage are passed to the second stage for further refinement. Then, the last stage is used to locate the final bounding boxes and landmarks as shown in Figure 4.8. MT-CNN is trained at a base learning rate of 0.001 and the learning

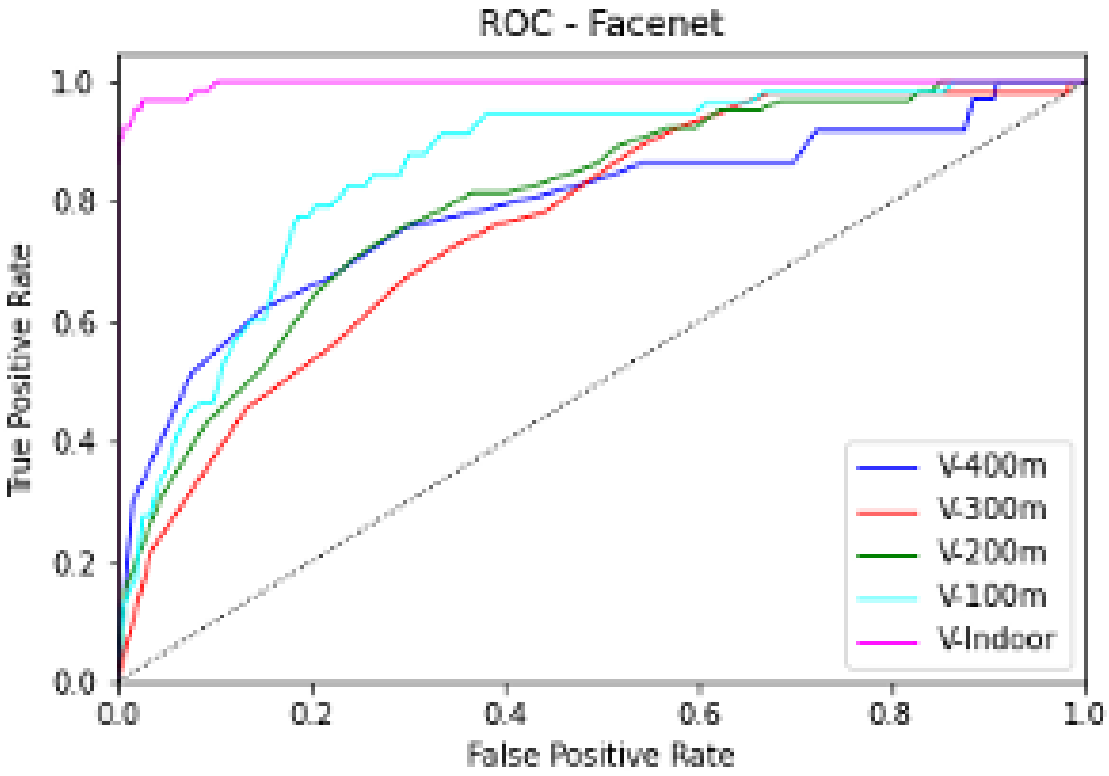

Figure 4.6: ROC curves for same spectral face recognition using Facenet.

rate is decreased linearly after 100 epochs until it reaches 0 at the end of 200 training epochs. A batch size of 16 is used to train the model. The baseline results obtained from MT-CNN are presented along with other results in Figure **??** and in Table 4.5. All the hyper-parameters are established through grid search.

4.2.2 Image Synthesis and Eye Center Detection

Once the data is prepared, a StarGAN2 model is trained using a learning rate of 0.00004 and a batch size of 1. The image synthesis model in Mokalla and Bourlai, 2021b is trained with a batch size of 8. Since the proposed method uses MT-CNN to detect the landmarks after each iteration, batch sizes higher than 1 resulted in the exhaustion of computational resources. Therefore, the model is trained to optimize the loss function given in Equation 3.25 for 120K iterations with a batch size of 1. The loss terms' weights are set to $\lambda_{sty} = 1$, $\lambda_{ds} = 1$, $\lambda_{cyc} = 1$, and $\lambda_{aln} = 3$. All the hyper parameters are established through grid search. The trained model is then used to generate visible band face images from thermal test images from 100 subjects.

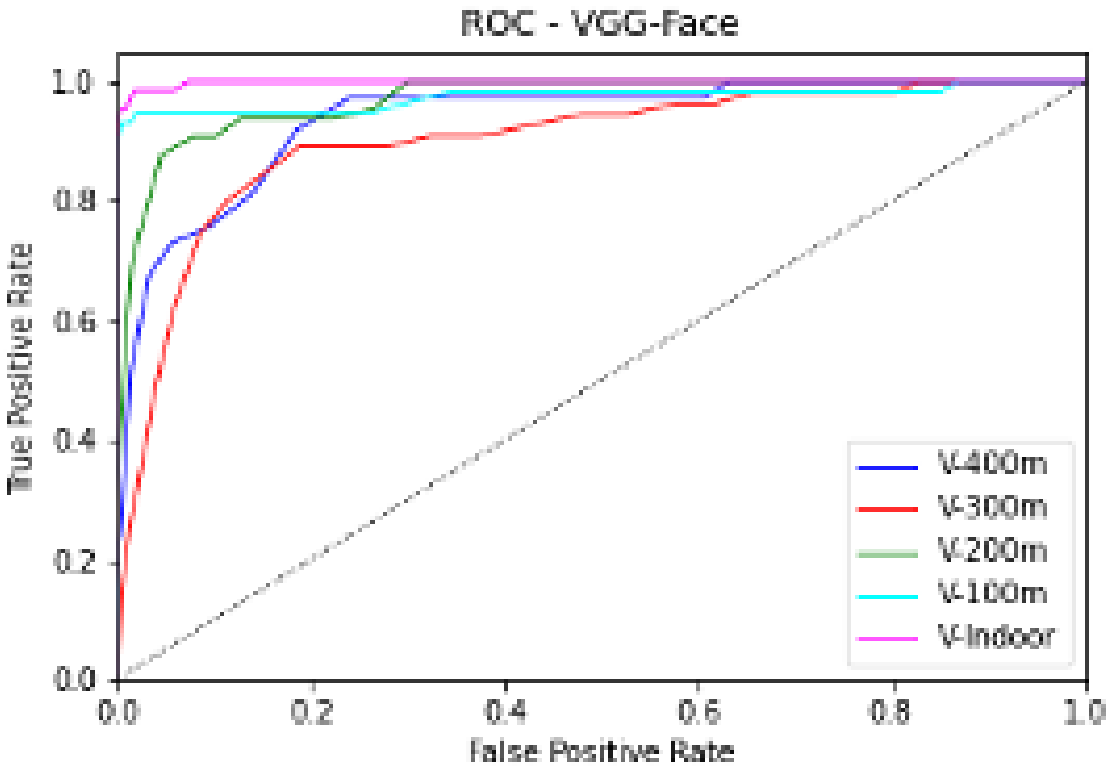

Figure 4.7: ROC curves for same spectral face recognition using VGG-Face.

CycleGAN is trained using an initial learning rate of 0.0001 for the first 100 epochs, then the learning rate is decayed linearly until it approaches zero at the end of 200 epochs. The batch size used to train this model without the alignment loss is 32, and with the additional loss term, as shown in Equation 3.15, it had to be reduced to 16 to not exhaust the computational resources. Similar to StarGAN2, all the hyperparameters are established through grid search, and $\lambda_{cyc} = 3$ and $\lambda_{aln} = 8$ are the weights used for the loss terms.

Then, MT-CNN and HR-Net are used to detect the eye centers in the synthesized visible face images. Next, these coordinates are mapped to their corresponding original thermal face images. Examples of original thermal and synthesized visible images are shown in Figure 4.9. The eye center detection accuracy is then calculated using the weakest eye estimation between the detected and ground-truth coordinates. The normalized error in 3.11 is used to calculate the accuracy. In the normalized error, $e \leq 0.25$ (or 25% of the inter-ocular distance) roughly corresponds to the width of the eye (i.e. corner to corner), $e \leq 0.10$ roughly corresponds to the diameter of the iris, and $e \leq 0.05$ roughly corresponds to the diameter of the

Table 4.4: Visible-to-visible same distance face verification metrics using VGG-Face. VGG-Face yielded the best performance at all the distances except indoor. This can be accounted to the larger input window size for VGG-Face (224×224), which is 4 times greater than that of ArcFace.

Distance	AUC ↑	EER ↓	TAR@1% ↑	TAR@5% ↑
Indoor	99.85	1.72	95.52	98.51
100 Meter	97.43	05.17	93.10	94.83
200 Meter	96.97	09.23	59.95	88.10
300 Meter	90.02	15.27	21.55	56.04
400 Meter	93.73	14.40	41.12	71.51

Table 4.5: Normalized Error vs. Accuracy - The proposed approach increases the eye detection accuracy over the baseline by around 14%, 50%, and 30% for e $\leq$ 0.05, 0.10, and 0.25 respectively. e$\leq$0.05 roughly corresponds to the diameter of the pupil, e$\leq$0.10 to the diameter of the iris,and e$\leq$0.25 to the width of the eye. SG - StarGAN2, CG - CycleGAN, L_{aln}83.36 - Alignment error (proposed), aug - augmentation with in-plane rotation, M - MT-CNN, H - HR-Net

Model	e $\leq$ 0.05 (%)	e $\leq$ 0.10 (%)	e $\leq$ 0.25 (%)
Baseline	7.2	34.0	69.0
SG(M)	24.0	87.2	97.3
SG(H)	21.0	85.6	99.6
CG(M)	1.7	5.2	9.5
CG(H)	0.5	2.3	5.0
SG(M)+ L_{aln}	**38.0**	**94.5**	**100**
SG(H)+ L_{aln}	34.0	92.0	100
CG(M)+ L_{aln}	22.3	52.8	86.0
CG(H)+ L_{aln}	19.5	45.1	81.0
SG(M)+ L_{aln} + aug	37.6	94.0	100
SG(H)+ L_{aln} + aug	34.0	90.1	99.5
CG(M)+ L_{aln} + aug	22.0	50.3	85.2
CG(H)+ L_{aln} + aug	18.8	45.0	80.6

pupil. All the results obtained, i.e., using the baseline method, training the original GAN models without the alignment error, and the proposed method are presented in Figures 4.10 and 4.11 and Table 4.5.

4.2.3 Face Recognition Experiments

The main purpose of automatic eye center detection is to assist in an FR system's ability to geometrically normalize a face. If the locations are detected incorrectly, it affects the geometric normalization, which in turn affects the FR system's performance. Therefore, to further understand the importance of accurate eye detection in FR, a series of face recognition experiments are performed using Facenet Schroff et al., 2015, ArcFace Deng et al., 2019, and VGG-Face Parkhi et al., 2015. For these experiments, one image for every subject is geometrically normalized using the ground-truth eye centers to ensure correctly aligned faces in

Table 4.6: Face identification and verification results using VGGFace. SG - StarGAN2, CG - CycleGAN, L_{aln} - Alignment error (proposed), aug - augmentation with in-plane rotation, M - MT-CNN, H - HR-Net. Regardless of the landmark detection and GAN models used, the proposed approach always yielded better results.

Model	**StarGAN2**			**CycleGAN**		
	Rank-1 ↑	AUC ↑	EER ↓	Rank-1 ↑	AUC ↑	EER ↓
Manual	100.0	98.03	01.76	100.0	98.03	01.76
No-alignment	48.48	83.36	27.21	48.48	83.36	27.21
Baseline	93.00	89.24	23.44	93.00	89.24	23.44
M	99.60	93.03	15.55	12.30	45.60	59.54
H	100.0	94.52	14.71	10.80	32.30	65.10
SG(M)+L_{aln}	**100.0**	**98.86**	**01.49**	**92.70**	**89.23**	**14.23**
H+ L_{aln}	100.0	98.77	4.42	89.23	88.49	11.86
M+ L_{aln} + aug	99.60	95.06	06.78	92.00	85.23	15.20
H+ L_{aln} + aug	98.33	95.00	09.50	85.08	85.14	10.37

the gallery set. The query set consists of face images that are geometrically normalized using the manual or baseline or synthesis methods or images cropped and resized without alignment. Face verification (1-to-1), and face identification (n-to-1) are performed using the aforementioned models. Face identification is presented as Rank-1 accuracy and Cumulative Match Characteristic (CMC) curves, and face verification metrics are presented as Area Under Curve (AUC) and Equal Error Rate (EER) and Receiver Operating Characteristic (ROC) curves. All the face recognition results are presented in Tables 4.6, 4.7 and 4.8. CMC and ROC curves using StarGAN2 for synthesis are illustrated in Figures 4.12, 4.13, and 4.14, and using CycleGAN in Figures 4.15, 4.16, and 4.17.

Table 4.7: Face identification and verification results using ArcFace. SG - StarGAN2, CG - CycleGAN, L_{aln} - Alignment error (proposed), aug - augmentation with in-plane rotation, M - MT-CNN, H - HR-Net. Regardless of the landmark detection and GAN models used, the proposed approach always yielded better results.

Model	**StarGAN2**			**CycleGAN**		
	Rank-1 ↑	AUC ↑	EER ↓	Rank-1 ↑	AUC ↑	EER ↓
Manual	100	92.41	11.95	100	92.41	11.95
No-alignment	48.65	59.36	46.05	48.65	59.36	46.05
Baseline	62.86	62.69	45.76	62.86	62.69	45.76
M	98.51	74.37	40.48	08.32	21.58	75.49
H	100.0	71.51	37.72	07.56	33.27	79.35
M+L_{aln}	**100.0**	**98.83**	**01.49**	**88.32**	**89.16**	**21.26**
H+ L_{aln}	100.0	81.31	25.97	86.75	85.37	19.96
M+ L_{aln} + aug	92.77	90.13	12.83	87.66	88.60	15.25
H+ L_{aln} + aug	93.56	92.77	09.85	81.45	79.00	18.21

Table 4.8: Face identification and verification results using Facenet. SG - StarGAN2, CG - CycleGAN, L_{aln} - Alignment error (proposed), aug - augmentation with in-plane rotation, M - MT-CNN, H - HR-Net. Regardless of the landmark detection and GAN models used, the proposed approach always yielded better results.

Model	**StarGAN2**			**CycleGAN**		
	Rank-1 ↑	AUC ↑	EER ↓	Rank-1 ↑	AUC ↑	EER ↓
Manual	94.00	97.18	10.26	94.00	97.18	10.26
No-alignment	50.67	84.24	21.27	50.67	84.24	21.27
Baseline	62.5	97.98	16.78	62.5	97.98	16.78
M	98.51	79.85	24.58	12.28	33.21	67.67
H	97.5	85.39	24.85	11.28	45.26	69.22
M+L_{aln}	**99.65**	**97.94**	**05.22**	**89.26**	**86.33**	**22.13**
H+ L_{aln}	100.0	91.48	14.8	87.15	85.17	24.23
M+ L_{aln} + aug	91.66	87.24	15.16	89.34	88.26	12.15
H+ L_{aln} + aug	94.87	90.27	11.28	85.17	86.20	16.23

4.2.4 Discussion

It can be seen from Table 4.5 that the proposed method improved the eye center detection accuracy by 30%, 60%, and 31% when e $\leq$ 0.05, 0.10, and 0.25 respectively when using HR-Net, and by 26%, 58%, and 31% respectively when using MT-CNN. The efficiency of the proposed approach is more evident when using CycleGAN. When the original CycleGAN is trained, it failed to keep the alignment of the images at the slightest resulting in very low eye detection accuracy. The eye detection accuracy when the proposed approach is applied to train CycleGAN increased by around 20%, and 50%, and 70% when e $\leq$ 0.05, 0.10, and 0.25 respectively. The reason for the poor performance of the original CycleGAN model is that it includes only two losses in training the model, namely, cyclic loss, and GAN loss. This resulted in synthesized images using the CycleGAN to be of poor quality, which lead to the failure in restoring the alignment of the face, thereby the eye center coordinates. StarGAN2, on the other hand, includes a total of four networks as explained in Section **??**, which explains the reasons for its superior performance. This, however, comes with a cost, i.e., longer training time, and higher memory requirements. To obtain optimal results, CycleGAN needs to be trained for around three hours, whereas, to train StarGAN2 on the same dataset using the same computer, it takes around 45 hours.

Tables 4.7, 4.8 and 4.6 demonstrate the importance of correctly aligning the face images in improving FR accuracy. When face images are simply cropped and not aligned, the performance is always poor regardless of the FR model used. Manually annotated eye centers yielded the highest performance metrics as expected. Baseline method performed better than the unaligned faces, however there is room for improvement. Synthesis methods yielded significantly higher performance metrics. When images are synthesized using the original StarGAN2 model without including the alignment loss, the Rank-1 accuracy improved by 15%, 8%, and 7% using Facenet, ArcFace, and VGG-Face respectively. The proposed method yielded Rank-1 accuracy of 100% using any landmark detection model and any FR model except for Facenet with

MT-CNN, in which case the Rank-1 accuracy was 99.65%. A similar pattern of improvement in AUC by around 10% from baseline to the proposed method. CycleGAN performed really poor on its own, and the accuracy improved by over 75% by incorporating the alignment loss. It can also be observed that the in-plane rotation did not affect the FR accuracy (for instance, VGG-Face FR accuracy using StarGAN with alignment loss decreased from 100% to 99.6% when the dataset is augmented with in-plane rotated images).

Similarly, when CycleGAN is used, the Rank-1 FR accuracy increased by around 80%, 72%, and 77% when using VGG-Face, ArcFace, and Facenet respectively. Also, AUC increased by over 50% irrespective of the FR model used. It is inferred that the models with higher eye center detection accuracy when $e \leq 0.10$ geometrically normalize the images well and thereby demonstrate superior FR performance.

4.3 Effects of Demographics and Photometric Normalization on Image Translation GANs for Cross-Spectral Face Recognition

This section explains the experimental protocol followed in detail. First, pix2pix network is used to synthesize visible images from thermal images using the original images without any photometric normalization techniques to establish a baseline for the training parameters. Then, StarGAN2 is trained with the original images to establish a baseline. Then, a few photometric normalization techniques are applied to the images and the experiments are repeated.

4.3.1 Thermal-to-Visible Image Synthesis with Pix2pix

Baseline Experiments

Since, pix2pix requires paired images, i.e., each image in the thermal band needs to correspond to exactly one image in the visible band, with almost overlapping object of interest, all the training, validation, and test images are geometrically normalized and paired. Then, the model is trained with a learning rate of 0.0002 for varying number of epochs. The number of epochs for the first experiment are 200, where the model is trained for the first 100 epochs at a constant learning rate of 0.0002, then the learning rate is decayed linearly for the next 100 epochs until it reached 0 at the end of training. The batch size used is 32, since this is the largest that could be used without running out of CUDA memory during training. Once the model is trained, it is evaluated on the validation set by using one ground-truth geometrically normalized visible image per subject as gallery and one generated visible image per subject as probe images. The gallery image remains the same each time a new model is trained unless a photometric normalization technique is applied to the ground-truth visible images. Then a learning rate of 0.00002 is used to train the model with the same number of epochs. The accuracy did not improve and the model did not reach the desired convergence with this lowered learning rate.

Then the model is trained with a learning rate of 0.0002 for 400 epochs, with fixed learning rate for the first 200 epochs and then the learning rate is decayed linearly for the next 200 epochs until it reached

zero at the end of training. This yielded the highest face verification AUC of 58.3% and the lowest EER of 44.8%. Then the learning rate is decreased and the model is trained for the same number of epochs, the accuracy did not improve with the lowered learning rate. Then the model is trained with a learning rate of 0.0002 for 600 epochs, where the first 300 epochs used the constant learning rate, and then the learning rate is decreased linearly until it reached zero at the end of the training. This decreased the face verification AUC by 12% and increased the EER by 8%. Therefore, the baseline hyper-parameters for pix2pix are fixed at the initial learning rate of 0.0002 for the first 200 epochs and then decayed linearly until it reached zero over the next 200 epochs. The ground-truth and the generated visible images are shown in Fig. 4.18.

4.3.2 Effects of Beard and Skin Color on Image Synthesis

The development and test sets used in this study includes people with facial hair: beard, and mustache. From Figure 4.19, it can be observed that the individuals with beard and mustache affected the synthesis process by generating blurred images, which in turn affected the face recognition accuracy. To understand these effects, all the images with bearded faces are removed from the gallery and probe sets.

Another factor that affected the performance of pix2pixGAN negatively is the skin color tone, specifically the model generated blurred images when the faces are of color (African American, Asian Indian) etc. To understand these effects and improve the face recognition accuracy further, these images are removed from gallery and probe of the test set. This further improved the rank-1 face identification accuracy by 4%, and the AUC increased by around 10%, and the EER decreased by around 10%. The original thermal and visible images and the generated visible images are shown in Fig 4.19. Number of subjects removed from the test set is 46 i.e., there are 46 subjects in the test set that had either a bearded face or are non-Caucasian.

4.3.3 Effects of Image Quality

Since, the beard and skin color degraded the performance of the pix2pix model, and the aforementioned experiments are carried out manually, an image quality metric is employed to automate the process. Two reference based image quality metrics, namely Structural Similarity Index Metric (SSIM), and Universal Image Quality (UIQ) are employed for this purpose. This filters out the images whose quality score falls under a certain preset threshold. The threshold is varied to understand the effects of these metrics on the face recognition accuracy. It is observed from these experiments that the number of images filtered out did not significantly vary with the threshold. It can be noticed from Table 4.9 that SSIM and UIQ resulted in the same accuracy metrics. This is because the two image similarity metrics filtered out the same images. Applying image quality metrics increased the rank-1 identification accuracy by 4.2% and AUC by 10.7% and reduced the EER by 10.2%, which is almost equal to the manual screening method.

4.3.4 Photometric Normalization

Different photometric normalization techniques are applied to the thermal and visible images to understand their effects on the image synthesis. The photometric normalization techniques applied are CLAHE,

Table 4.9: AUC, EER, and Rank-1 face recognition accuracy for pix2pix when all the images are included in the test set vs removing bearded and colored face images vs using SSIM and UIQ for image quality assessment.

Test data (e)	AUC (%)	EER (%)	Rank-1(%)
All images	58.3	44.8	3
Manual	68.6	34	7
SSIM Assessment	69	34.2	7.2
UIQ Assessment	69	34.2	7.2

Table 4.10: AUC, EER, and rank-1 face recognition accuracy for pix2pix using photometric normalization techniques.

PN technique	AUC (%)	EER (%)
Original	58.3	44.8
CLAHE	61.2	40.2
LBSSR	57.9	44.9
CLAHE-LBSSR	60.5	43.3

LBSSR, CHALE LBSSR. Each of these techniques are applied on the original images before any geometric normalization. Since the PN techniques can only be applied to gray-scale images, all the visible RGB images are converted to gray-scale prior to applying any of PN techniques. After PN is applied, then the images are geometrically normalized, and used for training and testing. The hyper-parameters obtained from the baseline experiments are once again used for training in this step, i.e., an initial learning rate of 0.0002 for 200 epochs and a linear decay until it reaches zero at the end of 400 epochs of training. Images after photometric and geometric normalization in the visible and thermal bands are shown in Fig 4.20.

Once the training is complete, the same testing method discussed above is followed for evaluating the trained models. All the accuracy metrics are enumerated in Table 4.10 and the ROC (Region Operating Characteristic) curves are illustrated in Figure 4.22 respectively. It can be noticed that the FR accuracy increased when CLAHE and CLAHE-LBSSR are applied to the images, however this is not a significant improvement. When LBSSR alone is applied to the images, the accuracy decreased.

4.3.5 Thermal-to-Visible Image Synthesis using StarGAN2

Baseline Experiments with StarGAN2

StarGAN2 does not require and does not use paired images, i.e., images in the thermal band need not correspond to images in the visible band. According to Choi et al., 2020, images need to be cropped and resized to 256×256 before training the model and the cropped images are shown in Fig 4.21. These cropped images are used to perform the first set of baseline experiments. The model is trained to a total of 100,000 steps with a learning rate of 0.0004 and a batch size of 4. Intermediate checkpoints are saved from 40,000 to 100,000 steps for every 10,000 steps. The model performed the best at 50,000 steps with

Table 4.11: AUC, EER, and Rank-1 face recognition accuracy for StarGAN2 using photometric normalization techniques.

PN technique (e)	AUC (%)	EER (%)	Rank-1(%)
Original	90.8	19.2	25
CLAHE	63.6	40.6	9
LBSSR	57.2	45.8	3
CLAHE-LBSSR	57.4	46.1	4

25% rank-1, and 90.8% AUC and 19% EER. Therefore, the rest of the StarGAN2 models in this study are trained using these hyper-parameters. Once the images are generated, faces are cropped out from the ground-truth and synthesized images using MTCNN (Multi-Task Cascaded CNN) (cite MT-CNN) to remove the non-face background information from the images. These cropped images are used for face recognition.

In the second set, once the images are generated using the trained model from above, eye centers are detected using MT-CNN and the images are geometrically normalized. These geometrically normalized images are used for face recognition experiments. This resulted in lower Rank-1 accuracy, lower AUC and lower EER, degrading the over-all recognition performance. In the third set, images are geometrically normalized prior training, and the resultant model is used to generate images. This still affected the performance negatively. The synthesized and ground-truth visible images are presented in Figure 4.21.

Effects of Photometric Normalization

Similar to the later sets of experiments using Pix2pix, three photometric normalization techniques are applied to the original thermal and visible images before training. Once again, CLAHE, SSR, and CLAHE-SSR are used. The third technique is where CLAHE is applied to the original images, followed by SSR technique. The photometric normalization techniques are applied to thermal and visible images in training, validation and test sets. After training, the models are used to synthesize photometrically normalized visible images from the test set images. The face detection tool used for the original visible images cannot be used for these images, because of the normalization. Since the photometrically normalized ground-truth images have the exact alignment as the original images, the bounding boxes from the previous set of experiments are used here to crop the faces. This was successful as the bounding boxes were notably accurate. Then, the cropped ground-truth and synthesized images are used for face matching experiments. The results are presented in Table 4.11, and the ROC curves are shown in Figure 4.22 respectively.

It can be observed from Table 4.11 that StarGAN2 performs the best when trained and tested on original visible and thermal images without any geometric or photometric normalization.

4.4 Involution GAN: Rethinking Architecture to Improve the Performance of Cross-Spectral Face Recognition

This section describes the experimental protocol followed and the results obtained in detail.

4.4.1 Baseline Experiments

ArcFace (Deng et al., 2019) is one of the most popular SOTA face recognition models available in the literature. To establish a baseline, ArcFace is trained using transfer learning by freezing the initial layers. The final three layers are retrained using the three training datasets (ARL-VTF, indoor, and outdoor). The ArcFace loss is given in the equation below:

$$L_{arcface} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s\cdot cos(\theta_{y_i}+m)}}{e^{s\cdot cos(\theta_{y_i}+m)} + \sum_{j\neq y_i} e^{s\cdot cos(\theta_j)}} \tag{4.1}$$

where N is the number of samples in the batch, s is the scale factor, θ_{y_i} is the angle between the feature vector and the y_i^{th} class, and θ_j is the angle between the feature vector and the j^{th} class ($j \neq y_i$). The learning rate to train this model is set to 1e-3, and the model is trained for 200 epochs with a batch size of 16. The parameters, s and m are set to be 30 and 0.5, respectively, in the original paper, however these are set to be 6 and 2, respectively, in this work. This is because the original ArcFace model is designed to classify millions of classes, while our training set has a maximum of 320 classes. During testing, the final fully connected layer is removed, and the feature embedding network is used to extract 512-D features from the test images. Then, the cosine distances between the embeddings from the gallery and query sets (visible is the gallery set, and thermal, query) is calculated which are then used to compute the face verification metrics. All the matching results are presented in Tables 4.12, 4.14, and 4.15 and Figures 4.24, 4.25, and 4.26. The results are presented in terms of AUC and EER following the most recent literature (Peri et al., 2021; Zhang et al., 2017) to be able to provide a comprehensible comparison.

4.4.2 InGAN implementation and experiments with ARL-VTF dataset

Several experiments are performed to understand the importance of implementing GAN architecture using involution blocks. All these experiments are performed using the ARL-VTF dataset to establish the hyperparameters. First set is to train the original StarGAN2 model without any alterations with a learning rate of 1e-4 and a batch size of 8 for 60K steps.

Then, InGAN is implemented with involution as the atomic operation to build its components. Additionally, cosine distance between the original and the synthesized visible face images is added as an additional loss term as described in Equation 3.25. This model is trained for 150K steps at a learning rate of 3e-6 with a batch size of 2. The weights λ_{sty}, λ_{cyc}, and λ_{id} are set to 1, 2, and 1 respectively. The learning rate is lowered since the size of the model increased with the incorporation of the pre-trained ArcFace

Table 4.12: Table showing the face verification results on the ARL-VTF dataset. It can be seen that the AUC of the proposed model is around 38% and 25% higher than the baseline and StarGAN2 models respectively.

Model	**Rank-1** (%) ↑	**AUC** (%) ↑	**EER** (%) ↓
Baseline	44.61	60.05	43.78
StarGAN2	67.27	73.72	34.72
StarGAN2+ArcFace	81.08	79.94	26.89
InGAN	81.53	88.05	18.56
InGAN+ArcFace (Proposed)	**94.82**	**98.08**	**5.17**

Table 4.13: Table showing the face verification results on the ARL-VTF dataset compared to the other SOTA literature. It can be seen that the AUC increased by at least 4% and the EER decreased by 7% over the most recent work.

Model	**AUC** (%) ↑	**EER** (%) ↓
Zhang et al., 2017	85.7	20.0
Di et al., 2021	87.0	18.1
Fondje et al., 2020	90.1	13.1
Peri et al., 2021	94.1	12.1
Proposed	**98.08**	**5.17**

Table 4.14: Table showing the face verification results on the MILAB-VTF(B) indoor dataset. It can be seen that the AUC of the proposed model is around 38% and 25% higher than the baseline and StarGAN2 models respectively.

Model	**Rank-1** (%) ↑	**AUC** (%) ↑	**EER** (%) ↓
Baseline	45.72	62.09	42.40
StarGAN2	56.89	74.70	32.30
StarGAN2+ArcFace	67.27	80.13	28.00
InGAN	74.54	87.96	19.30
Peri et al., 2021	-	76.30	30.60
InGAN+ArcFace (Proposed)	**81.5**	**93.70**	**15.40**

and computing cosine loss at each iteration. Batch size is also lowered because of the memory constraints when implementing this model. All the hyperparameters are set through grid search.

Figure 4.23 presents the original visible and thermal face images and the generated face images using the InGAN model proposed in this work. Table 4.13 presents a comparative analysis of the recent cross-spectral FR works that use ARL-VTF dataset. It can be seen that our model outperforms all the other models. Figure 4.24 presents the receiver operating characteristic (ROC) curves.

Table 4.15: Table showing the face verification results on the MILAB-VTF(B) outdoor dataset. It can be seen that the AUC of the proposed model is around 30% and 12% higher than the baseline and StarGAN2 models respectively.

Model	**Rank-1** (%) ↑	**AUC** (%) ↑	**EER** (%) ↓
Baseline	32.73	53.80	47.50
StarGAN2	45.94	69.90	36.70
StarGAN2+ArcFace	57.21	72.14	34.70
InGAN	66.76	76.35	31.18
Peri et al., 2021	-	75.50	31.10
InGAN+ArcFace (Proposed)	**72.97**	**85.40**	**20.60**

4.4.3 Experiments using the MILAB-VTF(B) dataset

Once all the hyperparameters are established using the ARL-VTF dataset, MILAB-VTF(B) is used to perform all the aforementioned experiments. These experiments include training ArcFace to establish a baseline, training StarGAN2, then training the InGAN model. The only other work that used the MILAB-VTF(B) dataset thus far is Peri et al., 2021, and these results are compared with theirs in Tables 4.14 and 4.15. Figure 4.23 shows the original thermal and visible images and the generated visible face images using the proposed InGAN model.

4.4.4 Discussion

Table 4.12 shows that using the involution operation to build the RedBlks improves the face verification accuracy significantly. It can be seen that the AUC increased by 13% when image synthesis model is used proving that synthesis models work better for cross-spectral face recognition in general. The accuracy further increased by around 6% when the pretrained ArcFace model is used to extract the feature embeddings to be used to compute the identity loss.

The proposed model, InGAN architecture, that uses involutional neural networks to build residual blocks improved the accuracy over the baseline by 28%, and over the StarGAN2 model by 15%. This, combined with the identity loss yielded an AUC of 98.08% and an EER of 5.17%. Using involution to build neural network models not only improves the accuracy but also improves the speed of training. This is because the filter parameters are initialized based on the pixel neighborhood, unlike the CNNs, where the parameters are initialized randomly. Since the parameters depend on the pixel values, involution incorporates self-attention. More details relating involution to self-attention can be found in D. Li et al., 2021. Table 4.13 compares our results with SOTA synthesis based thermal-to-visible face verification, and it can be seen that our model outperforms the recent works by around 4%-13%.

Tables 4.14 and 4.15 present the results on the more recent and more challenging MILAB-VTF(B) dataset, and it shows that the proposed methodology can be applied to other datasets, and it still improves the accuracy. In particular, the proposed model increases the AUC over the baseline and StarGAN2

models by around 30% and 13% respectively for the indoor dataset. A similar trend can be seen for the results obtained using the outdoor dataset, where the AUC increased by 30% and 12% over the baseline and StarGAN2 models respectively.

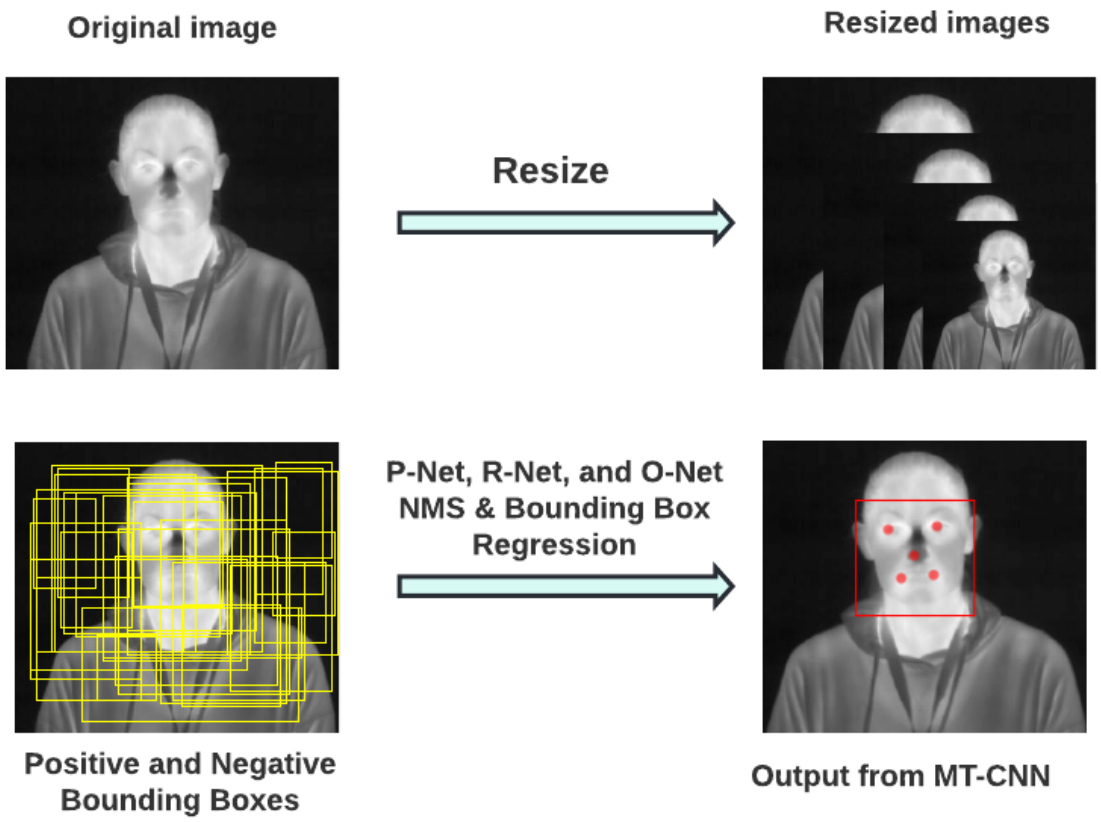

Figure 4.8: Figure showing the preparation of dataset and output from MTCNN. First, the images are resized to various scales, and each of the resized image is used for training along with the positive and negative parts. All the three stages of MTCNN (P-Net, O-Net, and R-Net) operate using bounding box regression and NMS.

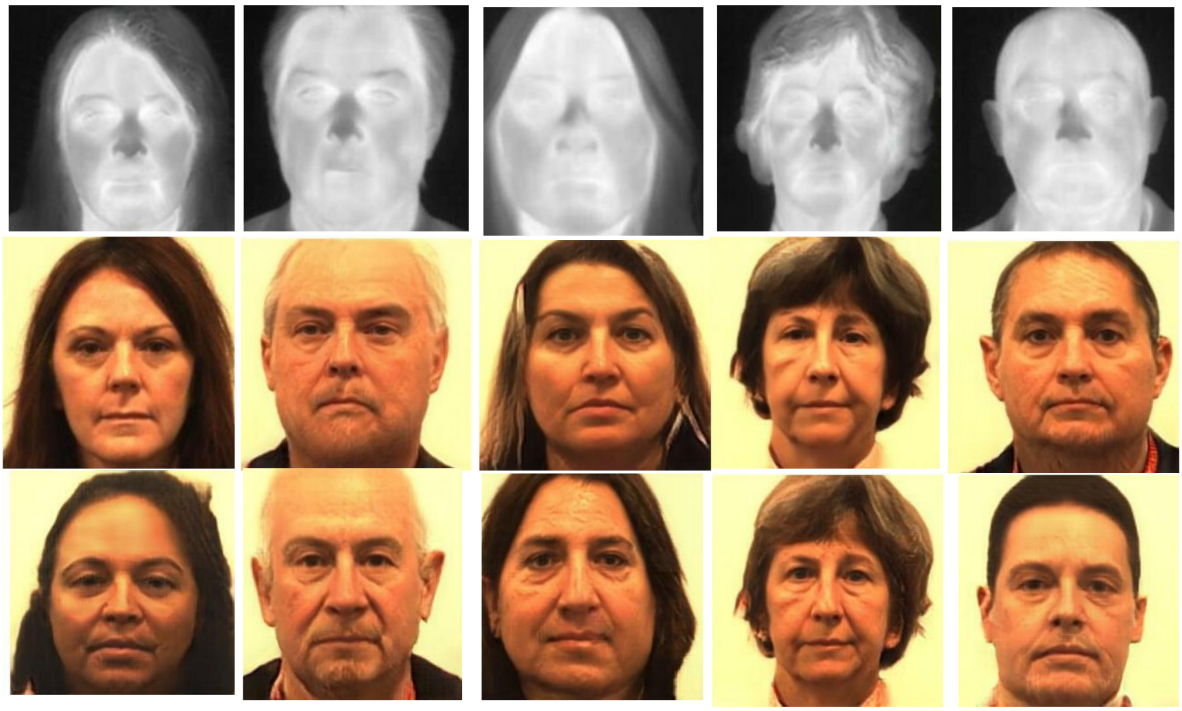

Figure 4.9: Examples of original thermal and synthesized visible images using StarGAN2 and CycleGAN. Eye centers are automatically detected in the synthesized visible images, and these are mapped to the original thermal images. First row shows the original thermal images, second and third row shows the synthesized visible images using the StarGAN2 and CycleGAN models trained with the proposed alignment loss, respectively.

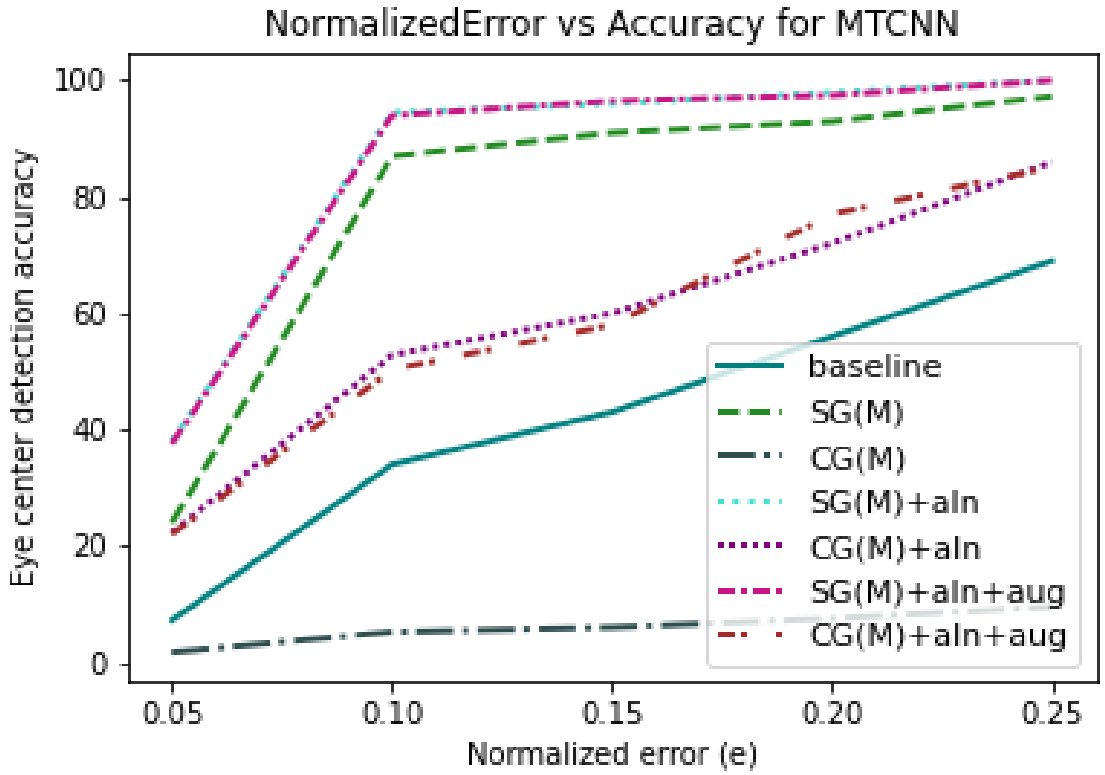

Figure 4.10: Normalized Error (e) vs Accuracy - The plot shows the improvement of eye center detection accuracy when using the proposed approach compared to the baseline and the original StarGAN2 and CycleGAN models. It shows an increase of around 30%, 60%, and 31% for e $\leq$ 0.05, 0.10, and 0.25 respectively, and by 26%, 58%, and 31%, respectively, when using MT-CNN. The eye center detection accuracy using CycleGAN improved by around 20%, 50%, and 70% when e $\leq$ 0.05, 0.10, 0.25 respectively.

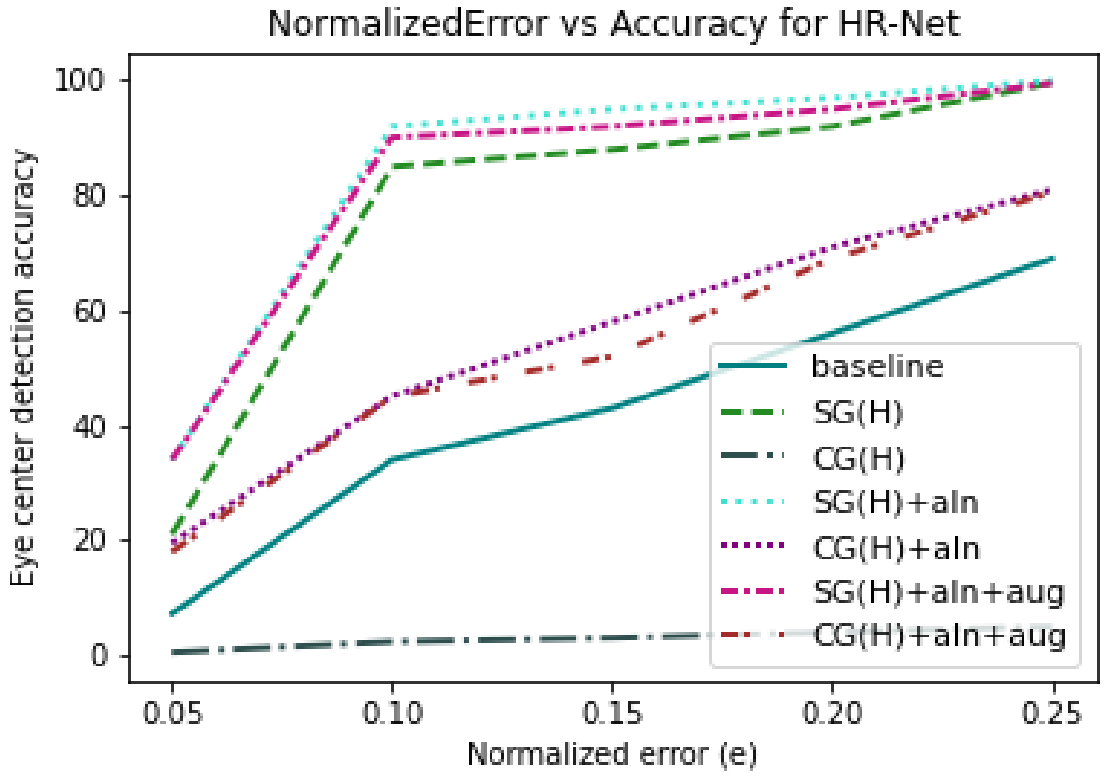

Figure 4.11: Normalized Error (e) vs Accuracy - The plot shows the improvement of eye center detection accuracy when using the proposed approach compared to the baseline and the original StarGAN2 and CycleGAN models. It shows an increase of around 30%, 60%, and 31% for e $\leq$ 0.05, 0.10, and 0.25 respectively, and by 26%, 58%, and 31%, respectively, when using HR-Net. The eye center detection accuracy using CycleGAN improved by around 20%, 50%, and 70% when e $\leq$ 0.05, 0.10, 0.25 respectively.

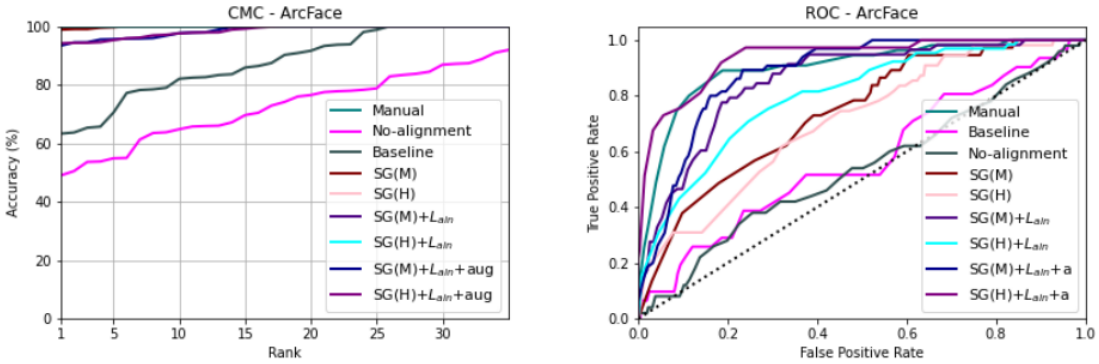

Figure 4.12: Face recognition plots when ArcFace is used with images synthesized using StarGAN2.

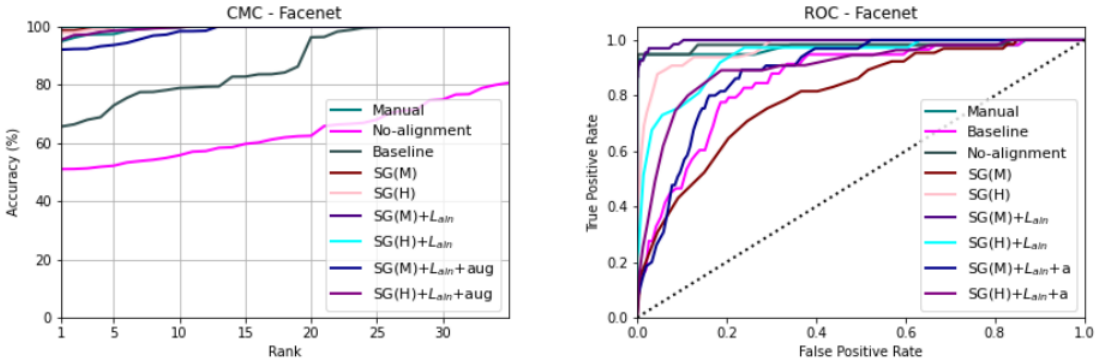

Figure 4.13: Face recognition plots when Facenet is used with images synthesized using StarGAN2.

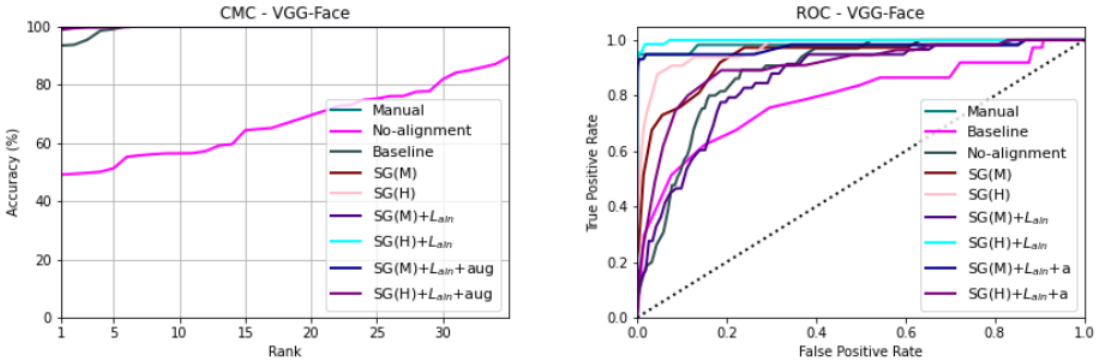

Figure 4.14: Face recognition plots when VGG-Face is used with images synthesized using StarGAN2.

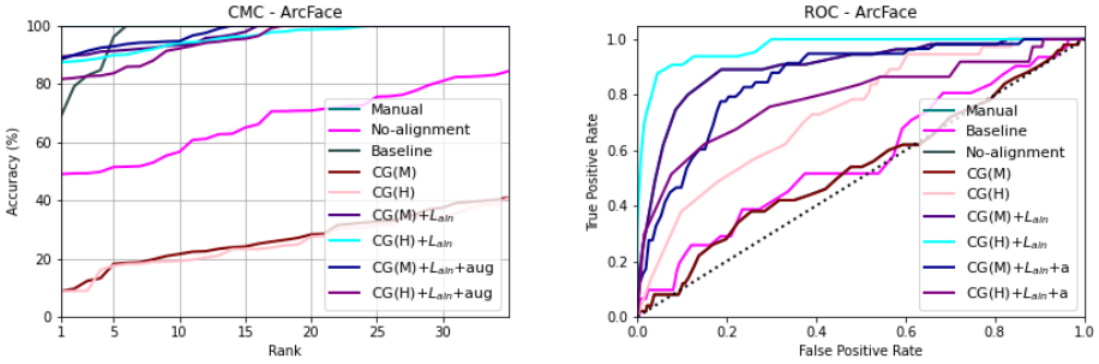

Figure 4.15: Face recognition plots when ArcFace is used with images synthesized using CycleGAN.

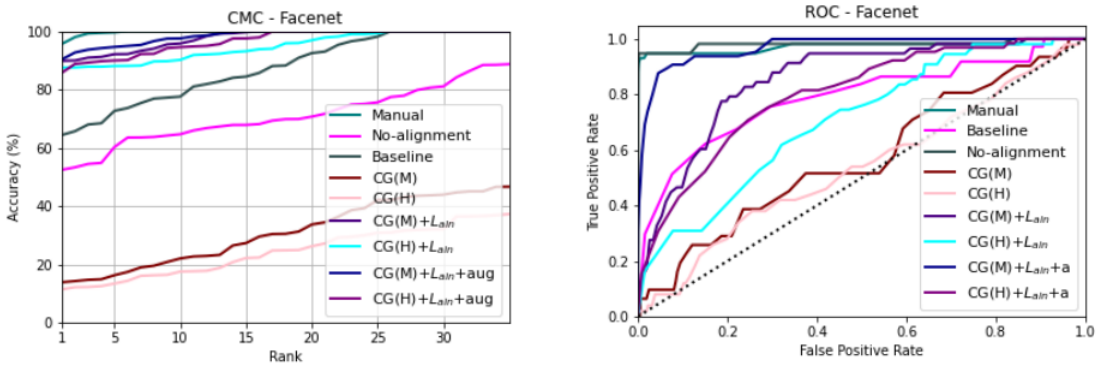

Figure 4.16: Face recognition plots when Facenet is used with images synthesized using CycleGAN.

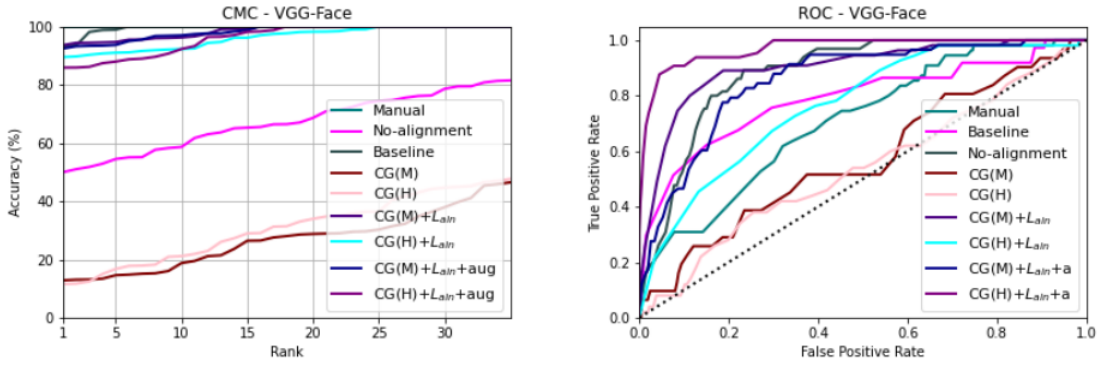

Figure 4.17: Face recognition plots when VGG-Face is used with images synthesized using CycleGAN.

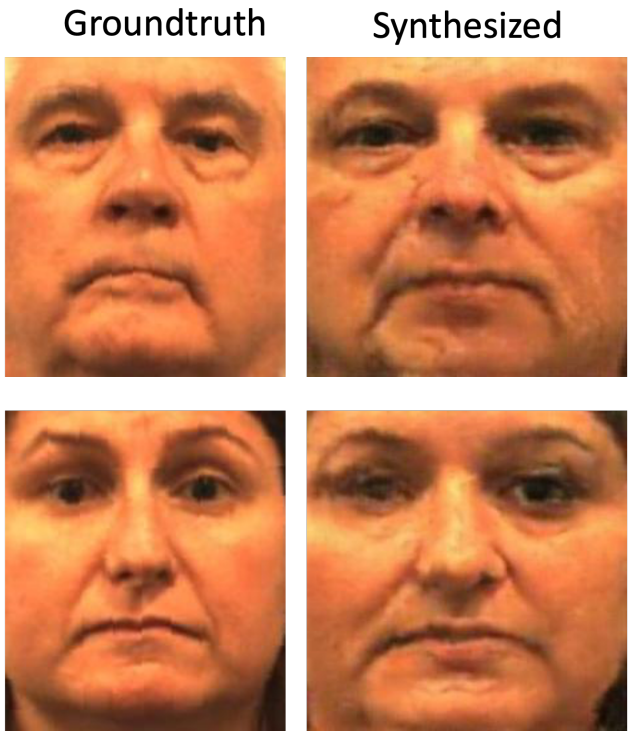

Figure 4.18: Original and synthesized visible images using pix2pix.

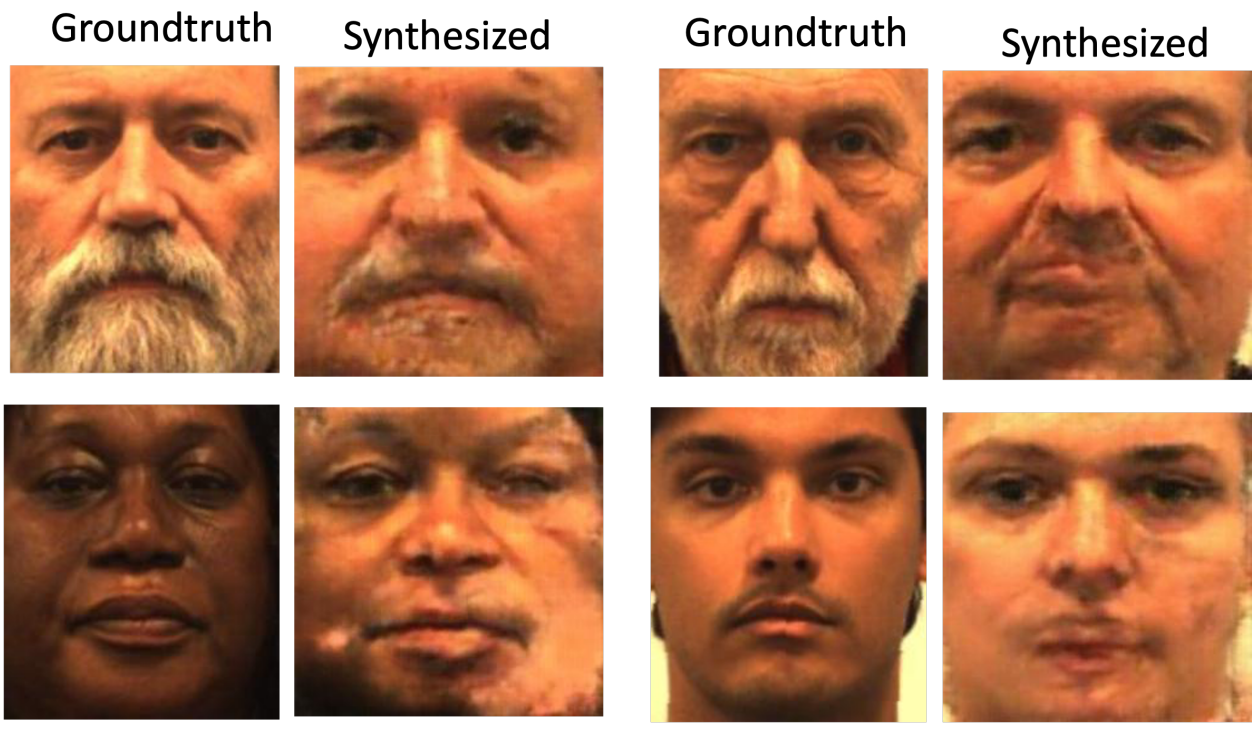

Figure 4.19: Synthesized images where the ground-truth images have bearded faces or non-Caucasian faces.

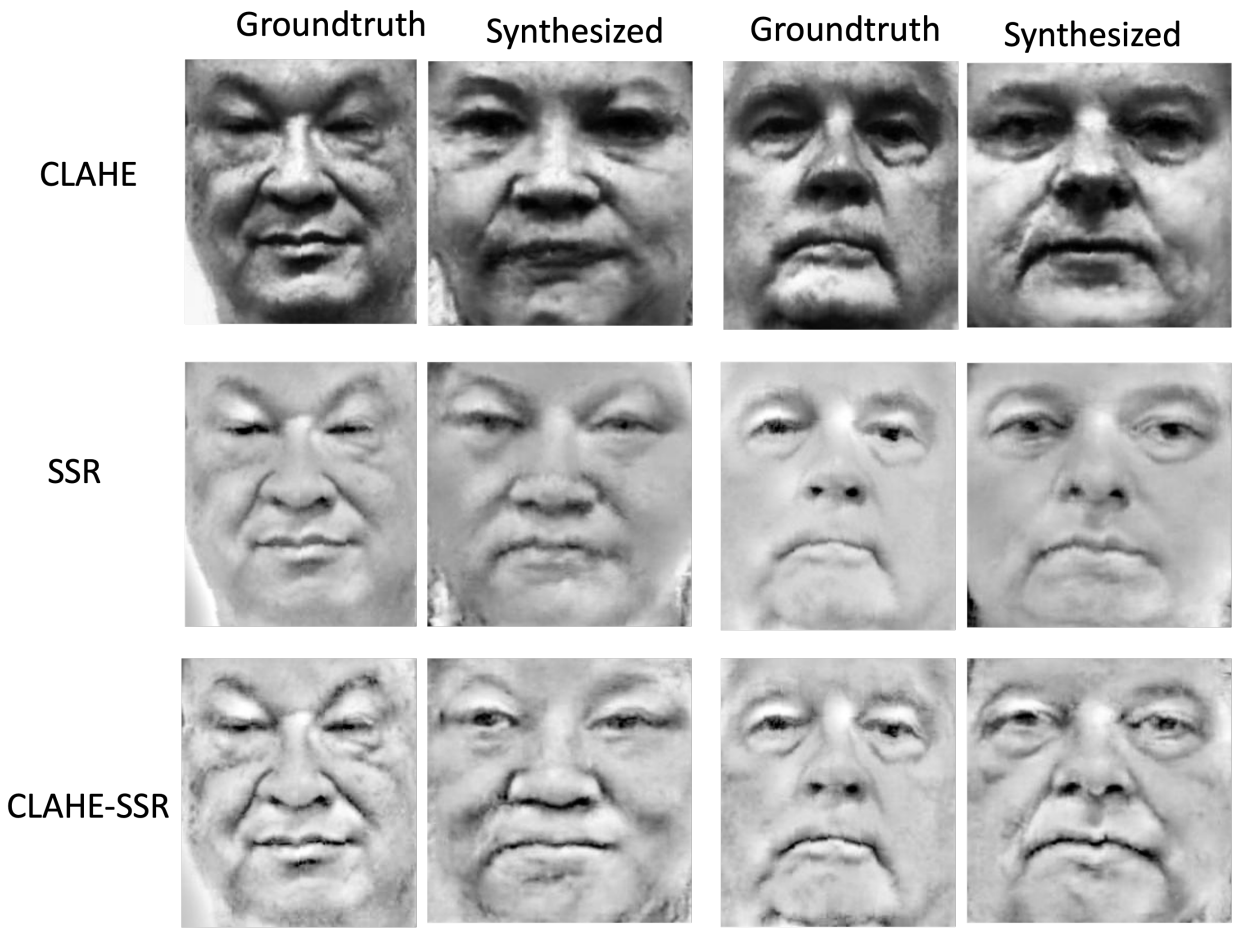

Figure 4.20: Ground-truth and synthesized visible images with PN techniques applied to the thermal and visible images before synthesis with Pix2pix.

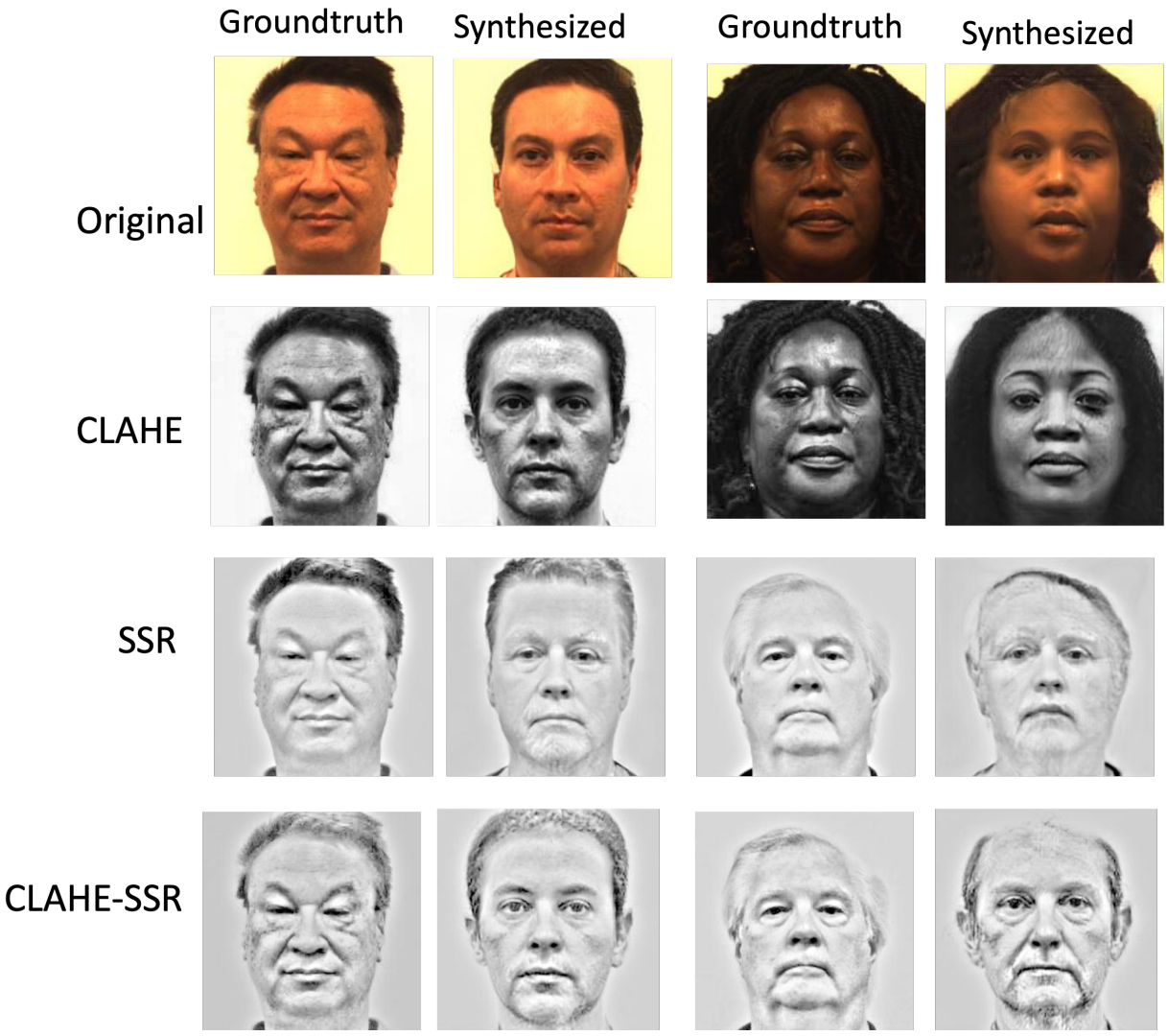

Figure 4.21: Ground-truth and synthesized visible images with PN techniques applied to the thermal and visible images before synthesis with StarGAN2.

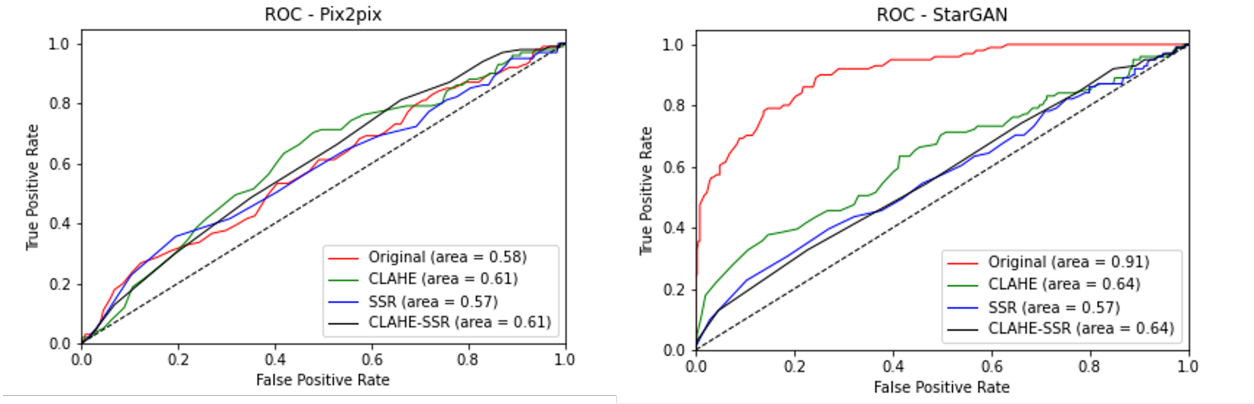

Figure 4.22: ROC curves for different PN techniques using pix2pix and StarGAN2 - Photometric Normalization techniques did not impact the performance of Pix2pix, however they affected the performance of StarGAN2 negatively.

Figure 4.23: Figure showing the original thermal and visible and the synthesized visible images from ARL-VTF, MILAB-VTF(B) indoor, and outdoor datasets. In each triad of images, the left image is the original thermal image, the second is the synthesized visible image, and the one on the right is the original visible image.

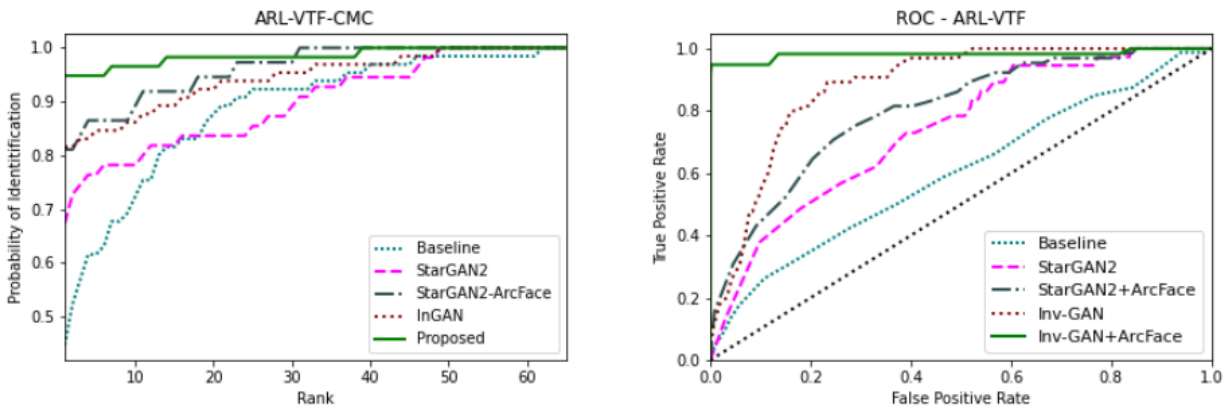

Figure 4.24: Figure showing the CMC and ROC curves for the ARL dataset. The curves shown in the figure prove the eminence of the proposed method over the baseline, and the original StarGAN2 methods. The proposed InGAN outperforms all the other methods irrespective of the dataset used.

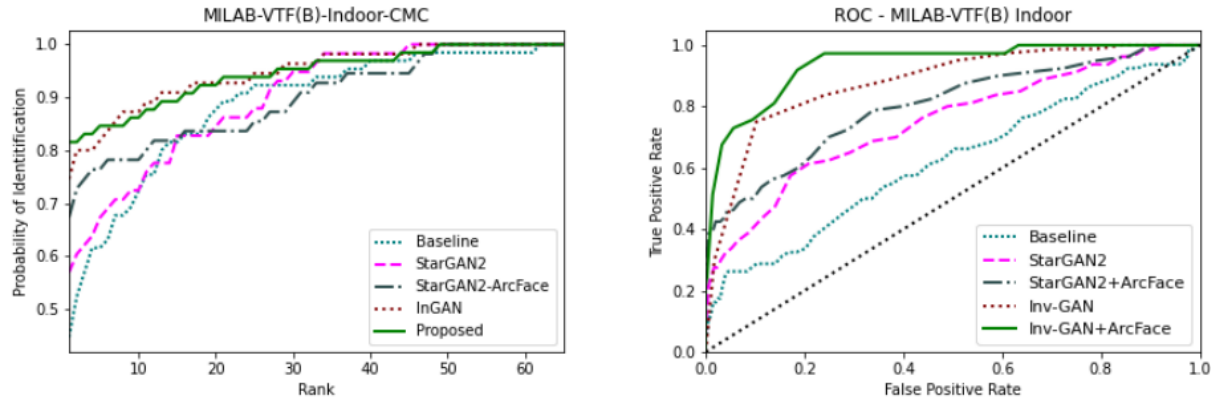

Figure 4.25: Figure showing the CMC and ROC curves for the MILAB-VTF(B) indoor dataset. The curves shown in the figure prove the eminence of the proposed method over the baseline, and the original StarGAN2 methods. The proposed InGAN outperforms all the other methods irrespective of the dataset used.

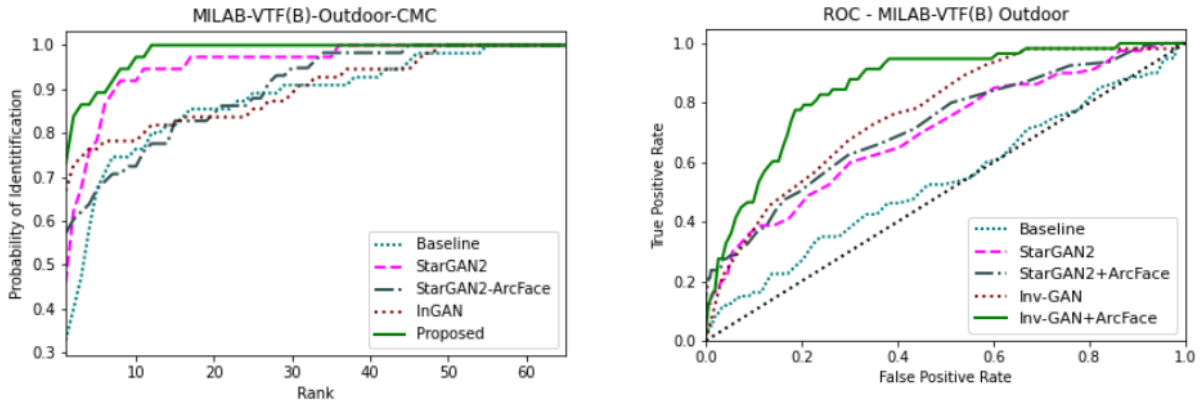

Figure 4.26: Figure showing the CMC and ROC curves for the MILAB-VTF(B) outdoor dataset. The curves shown in the figure prove the eminence of the proposed method over the baseline, and the original StarGAN2 methods. The proposed InGAN outperforms all the other methods irrespective of the dataset used.

Chapter 5

Conclusions and Future Work

This dissertation focuses on developing solutions for several problems in multi-spectral face recognition. There are several models that perform same-spectral face recognition, especially in the visible spectrum. Since the inception of convolutional neural networks, numerous models were proposed that work exceptionally well (Deng et al., 2019; Parkhi et al., 2015; Schroff et al., 2015). This is due to the availability of large-scale face datasets in the visible band (Guo et al., 2016; G. B. Huang et al., 2008). These datasets and the models proposed by training and evaluating on these datasets are not completely reliable concerning varying lighting conditions, such as, low-light to no-light scenarios. It is not as easy to train these models in the thermal spectrum due to the unavailability of many large-scale datasets in this band. This work addresses this problem using a synthesis approach. To address these problems, first a large-scale, mid-wave infrared (MWIR) and visible face dataset is collected, which is the largest yet of its kind, to the best of our knowledge. Then, an eye center detection solution in the thermal spectrum is proposed through image synthesis. Then, a series of experiments are conducted to understand the effects of demographics and photometric normalization on two image translation generative adversarial network (GAN) models. Finally, a framework is developed using involution (D. Li et al., 2021) as the atomic operation to build a generative adversarial network (InGAN) which is trained to synthesize visible images from their thermal counterparts. Specific conclusions are provided in the following sections.

5.1 Acquisition of Thermal-to-Visible Datasets

A new, large-scale face dataset of visible and MWIR thermal imagery is presented. Considering real-world applications and operational scenarios (where and how an operator will be able to utilize such a dataset), the dataset is structured to include head pose variations, as well as natural variations in expressions (uncontrolled environment), all captured using multiple visible and MWIR cameras, under indoor and outdoor conditions and at different stand-off distances, ranging from 1.5 meters to up to 400 metres. A curated version is in the works, and it is expected to be publicly available (restrictions apply).

5.1.1 Future Work

Currently, work is being done to create a curated version of the dataset, and it will be evaluated on a set of different tasks, including MWIR face, eye and ear detection, and recognition, MWIR facial landmark detection and same-spectral and cross-spectral matching, including identification and verification experiments using multiple state-of-the-art algorithms. There are other challenges that biometric researchers are expected to be working using this dataset, including cross-spectral face matching using severely degraded off-pose face images, or using subject wearing glasses (probe set) while they have been enrolled (gallery set) without any eye-wear. In the latter case and when operating in either the MWIR or LWIR bands, either eye correction glasses or sunglasses with result in facial occlusion induced by heat absorption in the lenses, and the expected consequence in face matching accuracy can be significant.

5.2 Utilizing Alignment Loss to Advance Eye Center Detection and Face Recognition in the LWIR Band

In this work, a synthesis based approach for eye center detection in the long-wave infrared (LWIR) spectrum is proposed. In addition to using StarGAN2 (Choi et al., 2020) and cycleGAN (Zhu et al., 2017) for image synthesis, *alignment loss* is incorporated, which is the normalized error between the predicted and actual eye center coordinates. To facilitate this, a pre-trained model is included in the StarGAN2 and cycleGAN training phases to extract the eye center coordinates from the generated visible images at the end of each iteration. The normalized error between these predicted eye center coordinates and the ground truth eye center coordinates is computed to constitute the alignment loss. This is used to train the models along with the adversarial and cyclic losses for cycleGAN, and along with the adversarial loss, style reconstruction loss, style diversification loss and cyclic loss for StarGAN2. There are many large-scale visible face datasets available with different types of annotations, facial bounding boxes, and facial landmarks, and reported results using both biometric functionalities, namely identification and verification. These wide variety of visible band face recognition related research led to the emergence of many state-of-the-art deep learning based models for face recognition, face detection, and facial landmark detection. However, visible datasets are not quite useful when operating in low light or nighttime conditions.

Using thermal face images is a viable solution however, the high cost of thermal imaging sensors, and the lack of large-scale thermal or multi-spectral datasets makes it difficult to train robust deep learning based models since the training of these models is data driven. To address this problem, visible images are synthesized from the thermal images and the available visible band face landmark detection models are used to detect the eye centers. These are later mapped to the original (corresponding) thermal face images. Results presented in 4.2 show the improvement of eye center detection accuracy of the proposed model over the baseline even in-plane rotated images are included. The aforementioned section also shows the improvement in face recognition accuracy when the proposed method is used over the baseline and the original GAN models. This proves the importance of accurately normalizing the face images to achieve better recognition performance.

5.2.1 Future Work

This work opens up the scope for future projects that can focus on expanding to other facial landmarks such as nose tip, mouth corners etc. We also intend to extend it to detecting more landmarks as resented in Kopaczka et al., 2016, which has applications in face tracking and pose estimation. This work addresses the problem eye center detection only in the frontal images because once the images are non-frontal, face alignment depends on other landmarks. This can be a potential project as well.

5.2.2 Effects of Demographics and Photometric Normalization on Image Translation GANs for Cross-Spectral Face Recognition

Since the thermal camera is invariant to ambient lighting conditions, face images collected in that spectrum can be used for nighttime surveillance. However, most of the legacy face datasets available are RGB images collected in the visible spectrum. Therefore, thermal-to-visible face recognition is of high importance for nighttime surveillance. This work focuses on two of the popular image-to-image translation GANs to synthesize visible images from their thermal counterparts and studies the effects of demographics and photometric normalization on the GAN models.

When using pix2pix, it is observed that the images with bearded faces and the images with non-Caucasian (colored) faces resulted in blurred outputs consistently. Removing these images from the test dataset improved the verification and identification accuracy. Two reference based image quality metrics, SSIM and UIQ are used to filter out the blurry images from the synthesized set. This improved the accuracy too. Then, CLAHE, LBSSR, and CLAHE-LBSSR are applied to the original thermal and visible images before training and testing the models.

StarGAN2 is not impacted by the presence of beard or by the ethnicity of the participant. However, the Rank-1 face identification accuracy is still lower than the acceptable range in practical applications. To understand the effects of photometric normalization on image synthesis, the aforementioned photometric normalization techniques are applied the original thermal and visible images.

5.2.3 Future Work

This work can be extended by applying many other photometric normalization techniques or combinations. The reason pix2pix failed in case of bearded and colored faces is because the data predominantly consists of Caucasian non-bearded faces. This problem can be addressed by augmenting the image data and training different models for each data type and creating a pipeline using beard/no-beard classification and ethnicity classification.

Since StarGAN2 performed the best for the given data, additional loss terms such as perceptual (Johnson et al., 2016), semantic (S. Liu et al., 2015) (Long et al., 2015) can be embedded into the GAN to improve the synthesis process, make the faces in the synthesized visible images look as close to the faces in ground-truth visible images as possible.

5.3 Involution GAN: Rethinking Architecture to Improve the Performance of Cross-Spectral Face Recognition

This study reimplements the GAN architecture using involutional neural networksimprove the thermal-to-visible face verification accuracy. Involution, introduced in D. Li et al., 2021, operates to invert the inherent properties of CNNs which are spatial-agnostic and channel-specific. Involution operation is spatial-specific and channel-agnostic and the filter weights are initialized based on the pixel neighborhood unlike the random initialization in CNN. This coheres with the fact that the different channels at one spatial position represent the same pixel, while different spatial locations represent different pixels. It also subsumes the self-attention since the initialization is not random.

These involutional neural networks are used as the building blocks for residual blocks and coin them *RedBlks*. GAN architecture consisting of three different blocks, namely, generator, discriminator, and style encoder are built using these *RedBlks*. The results show that the proposed method improves the face verification accuracy over the SOTA by 4% when using the ARL-VTF dataset, and by 17% and 10% with MILAB-VTF(B) indoor and outdoor sets respectively. Using involution not only increased the face verification accuracy, but also improved the training speed.

5.3.1 Future Work

This work opens up the scope to use involution to implement other neural network architectures to improve the system performance in other domains in terms of accuracy and training speed. This can also be extended to using the data collected at farther distances, for instance, MILAB-VTF(B) at 200, 300, and 400 meters by introducing face restoration (Abramian & Eklund, 2019; X. Wang et al., 2021), super-resolution (Johnson et al., 2016; T.-C. Wang et al., 2018) and other such abstractions.

Bibliography

Abaza, A., & Bourlai, T. (2013). On ear-based human identification in the mid-wave infrared spectrum. *Image and Vision Computing*, *31*(9), 640–648.

Abramian, D., & Eklund, A. (2019). Refacing: reconstructing anonymized facial features using GANs. *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019)*, 1104–1108.

Ahmad, N., Laskar, R. H., Husain, A., & Ahmed, M. (2020). Convolutional neural network based algoritm for eye center localization. *Journal of Natural Remedies*, *21*(7 (S1)), 1–5.

Anghelone, D., Chen, C., Ross, A., & Dantcheva, A. (2022). Beyond the visible: A survey on cross-spectral face recognition. *arXiv preprint arXiv:2201.04435*.

Asteriadis, S., Nikolaidis, N., Hajdu, A., & Pitas, I. (2006). An eye detection algorithm using pixel to edge information. *Int. symp. on control, commun. and sign. proc*.

Benamara, N. K., Zigh, E., Stambouli, T. B., & Keche, M. (2022). Towards a Robust Thermal-Visible Heterogeneous Face Recognition Approach Based on a Cycle Generative Adversarial Network. *Int. J. Interact. Multimed. Artif. Intell*.

Bourlai, T., & Jafri, Z. (2011). Eye detection in the middle-wave infrared spectrum: Towards recognition in the dark. *2011 IEEE International Workshop on Information Forensics and Security*, 1–6.

Bourlai, T., Narang, N., Cukic, B., & Hornak, L. (2012). On designing a SWIR multi-wavelength facial-based acquisition system. *Infrared technology and applications XXXVIII*, *8353*, 83530R.

Bourlai, T., Rose, J., Mokalla, S. R., Ananya, Z., Hornak, L., Nalty, C. B., Peri, N., Gleason, J., Castillo, C. D., Patel, V. M., & Chellappa, R. (2023). Data and Algorithms for End-to-End Thermal Spectrum Face Verification. *IEEE Transactions on Biometrics, Behavioral and Identity Science*, *2023*.

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). Vggface2: A dataset for recognizing faces across pose and age. *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 67–74.

Chaudhari, S. T., & Kale, A. (2010). Face normalization: Enhancing face recognition. *2010 3rd International Conference on Emerging Trends in Engineering and Technology*, 520–525.

Cheema, U., Ahmad, M., Han, D., & Moon, S. (2022). Heterogeneous visible-thermal and visible-infrared face recognition using cross-modality discriminator network and unit-class loss. *Computational Intelligence and Neuroscience*, *2022*.

Chen, C., & Ross, A. (2016). Matching thermal to visible face images using hidden factor analysis in a cascaded subspace learning framework. *Pattern Recognition Letters*, *72*, 25–32.

Chen, C., & Ross, A. (2019). Matching thermal to visible face images using a semantic-guided generative adversarial network. *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 1–8.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8789–8797.

Choi, Y., Uh, Y., Yoo, J., & Ha, J.-W. (2020). Stargan v2: Diverse image synthesis for multiple domains. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8188–8197.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, *1*, 886–893.

Deng, J., Guo, J., Liu, T., Gong, M., & Zafeiriou, S. (2020). Sub-center arcface: Boosting face recognition by large-scale noisy web faces. *European Conference on Computer Vision*, 741–757.

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Di, X., Riggan, B. S., Hu, S., Short, N. J., & Patel, V. M. (2021). Multi-scale thermal to visible face verification via attribute guided synthesis. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, *3*(2), 266–280.

Fondje, C. N., Hu, S., Short, N. J., & Riggan, B. S. (2020). Cross-domain identification for thermal-to-visible face recognition. *2020 IEEE International Joint Conference on Biometrics (IJCB)*, 1–9.

Fuhl, W., Santini, T., Kasneci, G., & Kasneci, E. (2016). Pupilnet: Convolutional neural networks for robust pupil detection. *arXiv preprint arXiv:1601.04902*.

Fuhl, W., Santini, T. C., Kübler, T., & Kasneci, E. (2016). Else: Ellipse selection for robust pupil detection in real-world environments. *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, 123–130.

Ghiass, R. S., Bendada, H., & Maldague, X. (2018). Université laval face motion and time-lapse video database (ul-fmtv). *Proceedings of the 14th International Conference on Quantitative Infrared Thermography*.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. *arXiv preprint arXiv:1406.2661*.

Guo, Y., Zhang, L., Hu, Y., He, X., & Gao, J. (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, 87–102.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, R., Li, Y., Wu, X., Song, L., Chai, Z., & Wei, X. (2021). Coupled adversarial learning for semi-supervised heterogeneous face recognition. *Pattern Recognition*, *110*, 107618.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hu, S., Choi, J., Chan, A. L., & Schwartz, W. R. (2015). Thermal-to-visible face recognition using partial least squares. *JOSA A*, *32*(3), 431–442.

Hu, S., Short, N. J., Riggan, B. S., Gordon, C., Gurton, K. P., Thielke, M., Gurram, P., & Chan, A. L. (2016). A polarimetric thermal database for face recognition research. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 119–126.

Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., et al. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7310–7311.

Huang, W., & Mariani, R. (2000). Face detection and precise eyes location. *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, *4*, 722–727.

Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. *Proceedings of the IEEE international conference on computer vision*, 1501–1510.

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

Jesorsky, O., Kirchberg, K. J., & Frischholz, R. W. (2001). Robust face detection using the hausdorff distance. *International conference on audio-and video-based biometric person authentication*, 90–95.

Jobson, D. J., Rahman, Z.-u., & Woodell, G. A. (1997). Properties and performance of a center/surround retinex. *IEEE transactions on image processing*, *6*(3), 451–462.

Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 694–711.

Kakadiaris, I. A., Passalis, G., Theoharis, T., Toderici, G., Konstantinidis, I., & Murtuza, N. (2005). Multimodal face recognition: Combination of geometry with physiological information. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, *2*, 1022–1029.

Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4401–4410.

Kevin, X., & Bowyer, W. (2003). Visible-light and infrared face recognition. *Workshop on Multimodal User Authentication*, 48.

Kopaczka, M., Acar, K., & Merhof, D. (2016). Robust facial landmark detection and face tracking in thermal infrared images using active appearance models. *VISIGRAPP (4: VISAPP)*, 150–158.

LeCun, Y., Bengio, Y., et al. (1995). Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, *3361*(10), 1995.

Li, D., Hu, J., Wang, C., Li, X., She, Q., Zhu, L., Zhang, T., & Chen, Q. (2021). Involution: Inverting the inherence of convolution for visual recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12321–12330.

Li, G., Cai, X., Li, X., & Liu, Y. (2006). An efficient face normalization algorithm based on eyes detection. *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 3843–3848.

Li, Y. (2019). Massface: An efficient implementation using triplet loss for face recognition. *arXiv preprint arXiv:1902.11007*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755.

Liu, S., Yang, J., Huang, C., & Yang, M.-H. (2015). Multi-objective convolutional learning for face labeling. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3451–3459.

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 212–220.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2018). Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, *15*(2018), 11.

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

Mallat, K., Damer, N., Boutros, F., Kuijper, A., & Dugelay, J.-L. (2019). Cross-spectrum thermal to visible face recognition based on cascaded image synthesis. *2019 International Conference on Biometrics (ICB)*, 1–8.

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Mokalla, S. R., & Bourlai, T. (2020). Face detection in MWIR spectrum. In *Securing social identity in mobile platforms* (pp. 145–158). Springer.

Mokalla, S. R., & Bourlai, T. (2021a). Effects of Demographics and Photometric Normalization on Image Translation GANs for Cross-Spectral Face Recognition. *2021 IEEE International Conference on Big Data (Big Data)*, 2109–2118.

Mokalla, S. R., & Bourlai, T. (2021b). Robust LWIR-based Eye Center Detection through Thermal to Visible Image Synthesis. *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, 1–8.

Mokalla, S. R., & Bourlai, T. (2023). Utilizing Alignment Loss to Advance Eye Center Detection and Face Recognition in the LWIR Band. *IEEE Transactions on Biometrics, Behavior, and Identity Science*.

Narang, N., & Bourlai, T. (2015). Face recognition in the SWIR band when using single sensor multi-wavelength imaging systems. *Image and Vision Computing*, *33*, 26–43.

Nguyen, D.-L., Putro, M. D., & Jo, K.-H. (2020). Human eye detector with light-weight and efficient convolutional neural network. *International Conference on Computational Collective Intelligence*, 186–196.

Osia, N., & Bourlai, T. (2014). A spectral independent approach for physiological and geometric based face recognition in the visible, middle-wave and long-wave infrared bands. *Image and Vision Computing*, *32*(11), 847–859.

Osia, N., & Bourlai, T. (2017). On Matching Visible to Passive Infrared Face Images Using Image Synthesis & Denoising. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 904–911.

Panetta, K., Wan, Q., Agaian, S., Rajeev, S., Kamath, S., Rajendran, R., Rao, S. P., Kaszowska, A., Taylor, H. A., Samani, A., et al. (2018). A comprehensive database for benchmarking imaging systems. *IEEE transactions on pattern analysis and machine intelligence*, *42*(3), 509–520.

Park, C. W., Kwak, J. M., Park, H., & Moon, Y. S. (2007). An effective method for eye detection based on texture information. *2007 International Conference on Convergence Information Technology (ICCIT 2007)*, 586–589.

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition.

Patel, H., & Upla, K. P. (2021). An unsupervised approach for thermal to visible image translation using autoencoder and generative adversarial network. *Machine Vision and Applications*, *32*(4), 1–18.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2536–2544.

Peri, N., Gleason, J., Castillo, C. D., Bourlai, T., Patel, V. M., & Chellappa, R. (2021). A synthesis-based approach for thermal-to-visible face verification. *2021 16th IEEE international conference on automatic face and gesture recognition (FG 2021)*, 01–08.

Poster, D., Thielke, M., Nguyen, R., Rajaraman, S., Di, X., Fondje, C. N., Patel, V. M., Short, N. J., Riggan, B. S., Nasrabadi, N. M., et al. (2021). A large-scale, time-synchronized visible and thermal face dataset. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1559–1568.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint*.

Reza, A. M. (2004). Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology*, *38*, 35–44.

Riggan, B. S., Short, N. J., & Hu, S. (2018). Thermal to visible synthesis of face images using multiple regions. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 30–38.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015:*

18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, 234–241.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Shafi, M., & Chung, P. (2009). A hybrid method for eyes detection in facial images.

Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5693–5703.

Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., & Wei, Y. (2020). Circle loss: A unified perspective of pair similarity optimization. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6398–6407.

Timm, F., & Barth, E. (2011). Accurate eye centre localisation by means of gradients. *Visapp*, *11*, 125–130.

Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*.

Valenti, R., & Gevers, T. (2011). Accurate eye center location through invariant isocentric patterns. *IEEE transactions on pattern analysis and machine intelligence*, *34*(9), 1785–1798.

Wang, H., Dong, X., Jin, Z., Dugelay, J.-L., & Tistarelli, M. (2021). Cross-spectrum Face Recognition Using Subspace Projection Hashing. *2020 25th International Conference on Pattern Recognition (ICPR)*, 615–622.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., & Liu, W. (2018). Cosface: Large margin cosine loss for deep face recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5265–5274.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al. (2020). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*.

Wang, P., Green, M. B., Ji, Q., & Wayman, J. (2005). Automatic eye detection and its validation. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, 164–164.

Wang, S., Liu, Z., Lv, S., Lv, Y., Wu, G., Peng, P., Chen, F., & Wang, X. (2010). A natural visible and infrared facial expression database for expression recognition and emotion inference. *IEEE Transactions on Multimedia*, *12*(7), 682–691.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., & Catanzaro, B. (2018). High-resolution image synthesis and semantic manipulation with conditional gans. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8798–8807.

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

Whitelam, C., & Bourlai, T. (2015). Accurate eye localization in the short waved infrared spectrum through summation range filters. *Computer Vision and Image Understanding*, *139*, 59–72.

Whitelam, C., Jafri, Z., & Bourlai, T. (2010). Multispectral eye detection: A preliminary study. *2010 20th International Conference on Pattern Recognition*, 209–212.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate analysis*, 391–420.

Wolf, L., Hassner, T., & Maoz, I. (2011). Face recognition in unconstrained videos with matched background similarity. *CVPR 2011*, 529–534.

Xiang, J., & Zhu, G. (2017). Joint face detection and facial expression recognition with mtcnn. *2017 4th international conference on information science and control engineering (ICISCE)*, 424–427.

Xie, S., & Tu, Z. (2015). Holistically-nested edge detection. *Proceedings of the IEEE international conference on computer vision*, 1395–1403.

Xingming, Z., & Huangyuan, Z. (2006). An illumination independent eye detection algorithm. *18th International Conference on Pattern Recognition (ICPR'06)*, *1*, 392–395.

Yadav, N. K., Singh, S. K., & Dubey, S. R. (2022). CSA-GAN: Cyclic synthesized attention guided generative adversarial network for face synthesis. *Applied Intelligence*, *52*(11), 12704–12723.

Yang, S., Luo, P., Loy, C.-C., & Tang, X. (2016). Wider face: A face detection benchmark. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5525–5533.

Yoo, D., Kim, N., Park, S., Paek, A. S., & Kweon, I. S. (2016). Pixel-level domain transfer. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, 517–532.

Yu, M., Tang, X., Lin, Y., Schmidt, D., Wang, X., Guo, Y., & Liang, B. (2018). An eye detection method based on convolutional neural networks and support vector machines. *Intelligent Data Analysis*, *22*(2).

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

Zhou, Y., & Berg, T. L. (2016). Learning temporal transformations from time-lapse videos. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, 262–277.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE international conference on computer vision*, 2223–2232.