Computer vision and deep learning applied to face recognition in the invisible spectrum

David Anghelone

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

THÈSE DE DOCTORAT

Vision par ordinateur et apprentissage profond appliqués à la reconnaissance faciale dans le spectre invisible

David ANGHELONE

Centre Inria d'Université Côte d'Azur, Équipe STARS
Thales DIS, Research Technology and Innovation

Soutenue le : 29 Juin 2023

Présentée en vue de l'obtention du grade de docteur en **Informatique** d'**Université Côte d'Azur**

Devant le jury, composé de :

Jean-Luc DUGELAY	Eurecom, France	Rapporteur
Christian RATHGEB	Darmstadt University of Applied Sciences, Allemagne	Rapporteur
Anil K. JAIN	Michigan State University, Etats-Unis	Examinateur
Touradj EBRAHIMI	Ecole Polytechnique Fédérale de Lausanne, Suisse	Examinateur
Marco LORENZI	Centre Inria d'Université Côte d'Azur, France	Examinateur
Antitza DANTCHEVA	Centre Inria d'Université Côte d'Azur, France	Directrice de thèse
Sandra CREMER	Thales DIS, France	Co-encadrante
Sarah LANNES	Thales DIS, France	Co-encadrante

Inria

THALES
Building a future we can all trust

Vision par ordinateur et apprentissage profond appliqués à la reconnaissance faciale dans le spectre invisible

Computer vision and deep learning applied to face recognition in the invisible spectrum

Jury :

Rapporteurs

Jean-Luc DUGELAY, Professeur, Eurecom, France

Christian RATHGEB, Professeur, Darmstadt University of Applied Sciences, Allemagne

Examinateurs

Anil K. JAIN, Professeur, Michigan State University, États-Unis

Touradj EBRAHIMI, Professeur, École Polytechnique Fédérale de Lausanne, Suisse

Marco LORENZI, HDR, Centre Inria d'Université Côte d'Azur, France

Antitza DANTCHEVA, HDR, Centre Inria d'Université Côte d'Azur, France

Sandra CREMER, Docteure, Thales, France

Sarah LANNES, Thales, France

"*Biometrics is about applications.*

Biometrics does not start with data and end with models/predictions. Rather, it starts with a problem faced by a real-world entity and ends with an action having an impact on that entity."

Courtesy, Karthik Nandakumar, MBZUAI

Résumé

La reconnaissance faciale cross-spectrale consiste à identifier des individus à partir d'images faciales provenant de différents spectres électromagnétiques, tels que l'infrarouge et le visible. Bien que cette application soit intrinsèquement plus difficile que la reconnaissance faciale classique, notamment en raison de la variation significative de l'apparence faciale causée par l'écart de modalité, elle est utile dans de nombreux scénarios impliquant la biométrie en vision nocturne ou la détection d'attaques par "présentation" des systèmes biométriques. Le but de cette thèse a été de développer un système de reconnaissance faciale thermique-visible, qui intègre de nouveaux algorithmes pour (a) la détection de visage thermique, ainsi que (b) la traduction d'image thermique en image visible, simplifiant le problème causé par l'écart de modalité pour la reconnaissance faciale cross-spectrale. Ainsi, tout algorithme peut être utilisé pour cette ultime étape de reconnaissance, un fait assuré par la plateforme de reconnaissance faciale FRP de Thales.

Pour atteindre cet objectif, nous présentons les contributions suivantes : Tout d'abord, nous avons collecté une base de données comprenant des images faciales multispectrales capturées simultanément sous quatre spectres électromagnétiques. La base de données fournit des ressources essentielles, riches et variées, qui sont appropriées pour reproduire des scénarios pratiques de fonctionnement d'un système biométrique cross-spectral. Deuxièmement, nous avons conçu TFLD, un algorithme de prétraitement dans la chaîne du système, qui consiste à détecter les visages et les points de repère faciaux dans le spectre thermique. TFLD est robuste à des conditions réalistes telles que la pose, l'expression, l'occultation, la mauvaise qualité d'image et la distance à longue portée. L'alignement facial résultant du prétraitement de TFLD a contribué de manière significative à l'amélioration des performances de reconnaissance faciale. La troisième contribution porte sur la traduction de spectres, visant à réduire l'écart de modalité entre le spectre visible et thermique pour la tâche de reconnaissance faciale. Nous avons présenté un nouveau modèle génératif guidé par l'espace latent, nommé LG-GAN, qui traduit un spectre (par exemple, thermique) en un autre (visible), tout en préservant l'identité d'un visage lors de la traduction. L'objectif de LG-GAN était lié à l'explicabilité, à savoir fournir un aperçu des caractéristiques biométriques pertinentes et discriminatoires à travers les spectres, qui a été poursuivi par l'élaboration d'un modèle génératif supplémentaire guidé par l'attention (AG-GAN). Enfin, pour répondre au défi de la reconnaissance faciale multi-résolution provenant de distances d'acquisition variables, nous avons proposé ANYRES. Ce modèle unique accepte toutes les résolutions d'images thermiques de visage et les convertit en images visibles synthétiques à haute résolution. De plus, nous avons étendu les capacités de ANYRES en mettant l'accent sur la reconnaissance faciale en environnement non contraint, considérant ainsi les variations de pose faciale.

Mots clés : Apprentissage Profond, Vision par Ordinateur, Intelligence Artificielle, Biométrie, Reconnaissance Faciale, Sécurité.

Abstract

Cross-spectral face recognition (CFR) refers to recognizing individuals using face images stemming from different spectral bands, such as infrared vs. visible. While CFR is inherently more challenging than classical face recognition due to significant variation in facial appearance caused by the modality gap, it is useful in many scenarios including night-vision biometrics and detecting presentation attacks. The aim of this thesis has been to develop an end-to-end thermal-to-visible face recognition system, which integrates new algorithms for (a) thermal face detection, as well as (b) thermal-to-visible spectrum translation, streamlined to bridge the modality gap. We note that any off-the-shelf facial recognition algorithm can be used for recognition, a fact ensured by the facial recognition platform (FRP) of Thales.

Towards the above goal, we present following contributions. Firstly, we collected a database comprising of multi-spectral face images captured simultaneously under four electromagnetic spectra. The database provides essential, rich and varied resources well-suited to replicate practical scenario of real-life operation for a cross-spectral biometrics system. Secondly, we proposed a preprocessing algorithm, streamlined to detect face and facial landmarks (TFLD) in the thermal spectrum. We designed the algorithm to be robust to adversarial conditions such as pose, expression, occlusion, poor image quality, long-range distance and related face alignment contributed significantly to improving face recognition scores. The third contribution had to do with spectrum translation, aimed at reducing the modality gap of visible and thermal spectrum w.r.t. face recognition. We presented a novel model, Latent-Guided Generative Adversarial Network (LG-GAN), which translates one spectrum (e.g., thermal) to another (visible), while preserving the identity across different spectral bands. The main focus of LG-GAN is related to explainability, namely providing insight in pertinent salient features, discriminative across spectra, has additionally been pursued by an additional model, namely the Attention-Guided Generative Network (AG-GAN). Finally, tackling the challenge of multi-scale face recognition stemming from varying acquisition distance, we proposed an algorithm, ANYRES, which accepts any resolution of thermal face images, proceeding to translate such into synthetic high-resolution visible images. We extended ANYRES to settings encountering users experiencing (extreme) facial poses, placing emphasis on thermal-to-visible face recognition in unconstrained environments.

Keywords: Deep Learning, Computer Vision, Artificial Intelligence, Biometrics, Face Recognition, Security.

Acknowledgements

It is with an immense pleasure and gratitude that I express my sincerest thanks and acknowledge to all the wonderful individuals who have contributed to the culmination of this dissertation. Their support and assistance have played a pivotal role in this journey.

I have received countless help and encouragements during these three years of Ph.D. I am immensely grateful to my advisors, Antitza, Sarah, and Sandra, for their exceptional guidance throughout my doctoral studies. Your expertise, mentorship, and patience have been instrumental in shaping my research and helping me to overcome challenges along the academic and industrial way. I am honored to had the opportunity to work under your supervisions.

Antitza, I extend my gratitude to you for sharing your expertise. I delved into the captivating realm of Biometrics, a field that I enthusiastically integrated with artificial intelligence. I am thankful for the opportunity and privilege you granted me to undertake this research, and thus this Ph.D.

Sarah, we have finished so many projects together and you provided me so much help in my career. I really enjoyed the time we wrote code and displayed images together. Your guidance, encouragement, and patience have been instrumental and you really pushed me to strive for excellence.

Sandra, thank you for giving me the opportunity to join your team that demonstrates excellence. This collaboration has offered me a valuable starter for my industrial career. I have also appreciated all your kindness in our interactions. Thank you for your valuable advice that will serve me for the rest of my life.

I would also like to express my gratitude to Sabine and Philippe for their supports during the initial years of my Ph.D. Without their guidance and encouragement, this journey would not have been possible. My sincere thanks also go to Monique for always being available and providing invaluable assistance whenever needed. Additionally, I would like to thank Francois for his valuable advice, which has not only contributed to my research but also had a positive impact on my personal growth. I would like to express my sincere appreciation to Sandrine, for her precious assistance and support throughout my research.

Furthermore, I would also like to thank my collaborator, Cunjian Chen, for the insightful discussions and the exchange of ideas that have enriched my research. Your contributions have played a significant role in shaping the outcome of this work.

I would like to extend my deepest appreciation to my esteemed jury members. It was an honor to have you as part of this important milestone. I am especially grateful to my thesis reviewers, Prof. Dugelay and Prof. Rathgeb, for their invaluable time and effort in reviewing my manuscript. Your insightful comments and suggestions have greatly enhanced the quality of my research. I am also grateful to my examiners Prof. Jain, Prof. Ebrahimi and Dr. Lorenzi for your insightful comments and constructive questions which have shaped my research.

I am grateful to ANRT for generously funding my research under the CIFRE-fellowship. I truly enjoyed my time at Thales and Inria, with all the fantastic members in RTI team, Anouar, Christine, Hania, Khanh, Amit, Thibaut ; and STARS team, Thibaud, Yaohui, Snehashis, Valeriya, Tomasz, Di, Hao, Aglind, Abid, Tanay, Mohammed, Rui, Srijan, Ujjwal, Michal, Laura, Rachid, Jean-Paul. We shared so much unforgettable moment ! This collaboration has enriched my research and provided me invaluable real-world insights.

Last but not the least, I am deeply grateful to Akim Falou Dine and my family for their support and love, and I would like to give special thanks to my girlfriend, Manon, for always being with me and encouraging me. You have been a constant source of motivation.

Contents

List of Figures

List of Tables

Dedicated to my grandparents, who I hope, are proud of me from the stars...

Chapter 1

Introduction

1.1 Context and Motivation

Facial recognition (FR) constitutes a biometric technology, aiming at *identifying* or *verifying* the identity of an individual by analyzing their distinct facial features [71, 69] and therefore at answering the questions 'who are you?' or 'are you the person you claim to be?'. Generally speaking, an automated FR-biometric system typically includes a sensor that captures biometric samples, *i.e.,* faces, and computer vision algorithms that process such samples, in order to extract relevant features for identification or verification. Specifically, the extracted features are compared to an existing dataset of similar features with the goal to determine whether a match exists. The dataset, referred to as *gallery*, contains the biometric template information, *i.e.,* the mathematical representation of unique biometric characteristics from individuals, who are already enrolled in the system. We denote the *probe* to be the biometric sample that is being compared to the gallery towards recognizing an individual.

FR has been a highly active research field for the last several decades [71, 153, 1, 92, 136]. Advances in deep convolutional neural networks (CNNs) and the associated seminal works DeepFace [129] and DeepID [126] have brought to the fore a remarkable progress in this area, tackling a number of challenges including variations in pose, illumination and expression (PIE), as well as resolution and unconstrained settings. While such works have predominantly focused on the visible spectrum, considering additional spectra allows for increased robustness [53, 13], in particular in the presence of variations of *pose* and *illumination*, as well as *noise* and *occlusions*. Further benefits include incorporating the *absolute size of objects*, as well as *robustness to presentation attacks* such as the use of makeup and masks to circumvent a FR-system [123].

Therefore, comparing RGB face images against face images acquired beyond the visible spectrum, often referred to as Cross-spectral Face Recognition (CFR), is of particular significance in designing FR systems for *defense, surveillance, and public safety* [53] and falls under the broader category of Heterogeneous Face Recognition (HFR). Here, *heterogeneous* indicates that face images have been acquired in different sensing modalities *e.g.,* visible (VIS), infrared (IR), 3D, and hand-drawn sketches (see Figure 1.1).

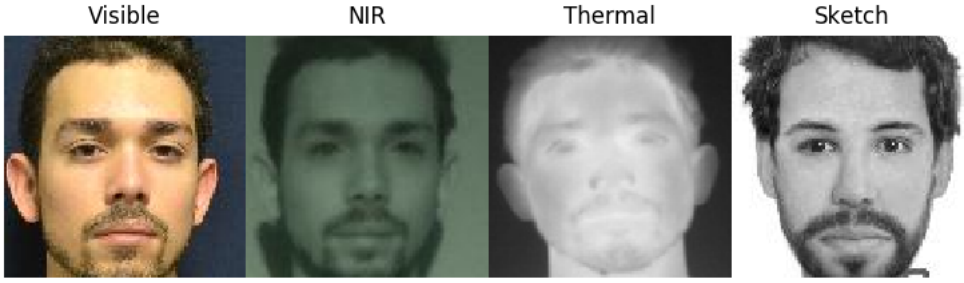


FIGURE 1.1: Examples of **heterogeneous face images** of the same subject captured in different modalities, *i.e.,* domains. Images from the Tufts dataset [110].

Beyond *access control* where acquisition is controlled *w.r.t.* user collaboration, the application of CFR has strong potential to provide new capabilities in *law enforcement*, *military* and *civil sector*, especially for commercial purposes [85]. Infrared-to-visible CFR has emerged as a promising solution, being able to tackle challenges encountered by traditional FR systems. The infrared (IR) spectrum refers to the range of electromagnetic radiation with wavelengths longer than those of visible lights, and can be divided into four sub-bands: near-infrared (NIR), short-wave infrared (SWIR), middle-wave infrared (MWIR) and long-wave infrared (LWIR). In this context, IR-sensors have several advantages over traditional visible light imaging, as related imaging can be employed for long-range monitoring and capturing biometric modalities at variable standoff distances, has the ability to capture images through certain types of obscuring materials, as well as in low-light or night-time scenarios. However, individuals who need to be identified in practical real-world situations often exhibit uncooperative behavior. The unconstrained environment in which identification is required, typically involves capturing images at a distance of subjects from the camera, accounting for facial poses or expressions, and even for occlusions. Therefore, developing methods that cater to real world unconstrained environments necessitate the tackling of such challenges.

1.2 Goals

This Ph.D. was funded by a french CIFRE-fellowship[1] supported by the Thales group [133]. The goal of this industrial thesis has been to develop algorithms towards an end-to-end thermal-to-visible CFR system, adaptive to unconstrained environments. Related algorithms range from *detection* to *recognition*, placing emphasis on an employment in a real-world surveillance scenario. Thermal denotes here the long-wave infrared (LWIR) band. We note that the choice for the spectral modality stems from the industrial need of Thales.

Towards this goal, in this thesis we have developed algorithms for (a) thermal face detection, as well as (b) thermal-to-visible spectrum translation for recognizing humans, with the main objective to function seamlessly with the

[1] Convention Industrielle de Formation par la REcherche

Thales's Face Recognition Platform (FRP) [38]. Below, we proceed to discuss our main goals.

(a) **Thermal Face Detection.** An initial step in identifying individuals in an unconstrained environment involves detecting their faces in thermal spectrum images. Biometric systems conventionally accept standardized images as input. Hence, aligning the localized faces based on normalized canonical coordinates provides standardized *facial images*, that could be fed into the system.

We therefore focus on implementing a method of thermal face and facial landmarks detection tailored to be robust in unconstrained environment and able to provide facial key-point contributing to improve FR scores. This facial detector serves as a preprocessing step of facial alignment images, thus facilitating identification.

(b) **Thermal-to-visible spectrum translation for recognizing humans.** Towards comparing cross-spectral faces, we aim at finding representations that allow for extraction of robust biometric features beyond the visible spectrum. Specifically, spectrum *translation* methods relying on generative adversarial networks (GANs) are of particular pertinence in allowing for reduction of the modality gap. In this context, a facial image is translated from one spectrum (*e.g.,* thermal) to another (visible), thereby creating a fully artificial facial visible-like image, which serves as the "probe-image" that is being compared to the gallery of visible reference face images. Spectrum translation is a type of image-to-image translation, which is a broad class of computer vision tasks aimed at transforming an image from one domain to another. In this process, a generative model learns to map an input image to an output image, typically by minimizing a loss functions that measure the difference between the generated and target images.

We control facial image generation ranging from pixels to features level, in developing instrumental loss functions that enable the generator to produce high-fidelity (realistic) visible-like face image, as well as constraining discriminative identity feature preservation across spectra.

Face is considered as one of the prominent biometrics, as it can be acquired in unconstrained settings, without the consent of an individual. In designing an operational CFR system that is adaptive to unconstrained environments, we place emphasis on *uncontrolled recognition*. Specifically, real-world CFR scenarios entail FR at random *distance* from the camera, thus resulting in different (uncontrolled) image resolutions, where *pose changes* are additional factors that the system should be robust to.

We have aspired to design a reliable automated generative model, able to accept any (low) resolution thermal face image as input, and to translate it into a visible-like realistic face image in higher resolution. Extending the goal of preserving faithful biometric features across *spectra*, as well

as *resolution*, emphasis is placed on building a unique model which handles simultaneously dual computer vision tasks, *i.e.,* super-resolution and domain translation.

While addressing above enlisted goals, we identified a number of challenges that we proceed to present in the next Section.

1.3 Challenges

CFR is more challenging than traditional FR for both *human examiners* as well as *computer vision algorithms* [53]. Particularly, the task of CFR becomes even more arduous when scenarios of application involve recognizing individuals in unconstrained environments. This thesis is addressing key issues related to thermal-to-visible CFR and we have identified a set of factors increasing challenges of operational CFR system that we proceed to enlist.

Large intra-class variability. Human faces exhibit large intra-class variation, within the same spectral modality, where face samples of the same subject can exhibit larger appearance variation than face samples of different subjects. This intra-class variation impedes the task of learning discriminative large-margin identity-related deep features, rendering accurate distinguishing between different individuals across spectra challenging, which is critical for reliable CFR. We have tackled this challenge in Chapter 5.

Spectral modality gap and identity preservation. CFR aims to compare faces sensed in different spectra, thereby comparing facial features emerging from two domains. In this context, the *modality gap* is of concern, where appearance variation between two face samples of the same subject can be larger than that of two samples belonging to two different subjects (see Figure 1.1). This can result in degraded FR performance and there is a need to address this large discrepancies. Therefore, *preserving the identity* of the same person across spectra is considered as a primary challenge, which we address in Chapter 5.

Insufficient data. Unlike traditional FR, which benefits from millions of face images available for training deep learning models, CFR is still hindered by the scarcity of large-scale datasets that encompass different spectral modalities. Furthermore, only a handful of existing datasets provide annotations of face bounding boxes and landmarks, which are crucial for developing automated face and landmark detection for CFR. Therefore, the *limited* availability of *training samples* of cross-modality face image pairs challenges the design of effective CFR methods using deep neural networks. Towards mitigating this challenge, we have collected a novel dataset, which we present in Chapter 3.

Poor image quality. Thermal imaging technologies typically offer low resolution thermal images that lack detailed facial information. Poor image quality, associated to the inherent poor texture information, low-contrast and low resolution, all impede the determination of salient facial discriminative identity

features, and thereby posing significant challenges for related CFR. We tackle this challenge in Chapters 4 and 6.

CFR in unconstrained environments. Thermal-to-visible CFR in unconstrained scenarios is highly challenging due to *e.g.,* uncontrollable environment effects. Unlike the cooperative recognition, *i.e.,* in which the user actively participates in the recognition by adopting correct and static postures in a constrained setting, the uncontrolled environment inherently incorporates dynamic phenomena. Individuals do not facilitate the image acquisition process, and hence factors such as recognition in day or night setting, as well as facial expression and poses impede FR. We tackle this challenge in Chapters 4 and 6.

CFR at a distance. Settings, where humans are at a random distance from the camera, bring to the fore multi-scale resolution images (as a function of acquisition distance). Hence, faces might be represented by a few pixels, containing limited biometric information, hindering facial detection, as well as facial recognition. Naturally, as image quality decreases, detecting facial landmarks that serve as fiducial points for facial image alignment (a preprocessing step for a FR-system), becomes a more challenging task in the context of thermal images. This can lead to inaccurate face alignment, which in turn can significantly impact the accuracy of FR system. We have addressed this in Chapters 4 and 6.

Operational CFR system. In order to design an operational CFR system, a facial detector is a crucial preprocessing step to localize an individual. Once faces are detected, it enables the alignment of images based on related detected landmarks for the subsequent facial feature extraction step. However, thermal imaging provides limited detail and presents challenges in semantically defining certain landmarks, especially in areas with significant heat diffusion, which makes accurate detection difficult. This results in an additional challenge for CFR. Gallery images are acquired in controlled settings with a high quality camera. However, thermal probe images provided by prior detector that can be of random low resolution are to be compared with high resolution gallery images. We note that traditional CFR-systems accept images of fixed resolution as input, which would require a specific model for each input resolution. This is completely impractical in real-life scenarios. Therefore, a unique model recovering accurate high-resolution visible face images from any (low) spatial resolution thermal aligned face images, while *preserving the identity* across both spectrum and resolution is critical in the design of operational CFR system. We elaborate and tackle this in Chapter 6.

All these described challenges influence the performance of a CFR system. In the next Section, we present our contributions aimed at tackling these challenges.

1.4 Contributions

Our contributions are motivated by the complex challenge involved in designing an end-to-end CFR system. We initially explored spectral bands suitable for CFR and examined techniques that specifically functioned in the infrared spectrum. In particular, we point out the need for multi-spectral resources, in which the lack of large-scale multi-spectral paired-face datasets inherently impedes CFR research. This comprehensive study resulted in the conclusion that the NIR-to-visible CFR scenario has been extensively studied in the literature, achieving high accuracy. However, the SWIR and MWIR sub-bands remain widely unused due to their impracticality and high equipment costs. We note that the COVID-19 pandemic has brought a significant decrease in the cost of LWIR sensors, rendering such sensors more accessible for a wider range of applications. This, coupled with the increasing demand for surveillance and protection solutions, *e.g.,* against spoofing attacks, has made LWIR imaging an appealing option for various industrials, including Thales.

Motivated by the above, we focus in this thesis on recognizing individuals with thermal sensors, namely in the LWIR-to-visible CFR scenario in *unconstrained* environments. We here note that all CFR implemented methods that we present are adaptable to other IR sub-spectra.

Major contributions

1. **Automatic face and landmark detection as a key preprocessing step in cross-spectral face recognition.** Towards proposing a thermal face detector, instrumental in automated thermal-to-visible CFR system, we introduce an innovative *thermal face and landmarks detector* (TFLD) method streamlined to be robust to adversarial conditions such as pose, expression, occlusion, poor image quality and long-range distance. Unlike existing methods, where facial landmarks denote specific marks, TFLD considers them as the center of a textured area. By adopting such strategy, TFLD detects pertinent facial key points for the purpose of face alignment, demonstrating a positive impact on related FR scores.

2. **Explainability as a guidance for cross-spectral identity discriminative feature preservation through spectrum translation.** We investigate *finding representations* that allow for extracting robust biometric features beyond the visible spectrum. Recent advances in artificial intelligence related to generative models present several advantages in the context of thermal-to-visible CFR. Hence, *spectrum translation* algorithms relying on GANs are of particular significance in allowing for reduction of the modality gap. Building on innovative deep learning that translates one spectrum (*e.g.,* thermal) to another (visible), while preserving *identity* across different spectral bands is the challenge that we address through two explainable models, namely, Latent-Guided Generative Adversarial Network (LG-GAN) and Attention-Guided Generative Network

(AG-GAN). The former explicitly decomposes a facial image into an identity code and a style code, where the identity code is learned to encode spectral invariant identity features between thermal and visible domains. In addition, the identity code offers useful insights to explain salient facial structures that are essential to the synthesis of high-fidelity spectrum face images. The latter involves attention feature maps to deepen knowledge on cross-spectral facial features that are learned during spectrum translation. In particular, visualization on the attention maps entails transparency and interpretability regarding salient facial features that are discriminative across spectra. Finally, a commonality between LG-GAN and AG-GAN has to do with the use of specific loss functions that are critical to the faithful identity preservation, as well as the ability to generate realistic faces, thereby enabling the development of more accurate and robust CFR systems.

3. **Adaptation for unconstrained CFR.** Although LG-GAN and AG-GAN have demonstrated their potential to preserve consistent biometric features across spectra, there is a need to adapt the system to unconstrained environment. We present ANYRES, a novel algorithm, designed to address simultaneously *facial super resolution*, as well as *spectrum translation*. Particular effort has been placed on being robust to a wide range of (low) resolution thermal image inputs, while preserving the identity information. During acquisition, humans are situated at a random distance from the camera, resulting in multi-scale resolution images, depending on the acquisition distance. ANYRES accepts therefore any resolution of thermal face images, which are translated into a synthetic high-resolution visible-like image. Additionally, ANYRES involves the tackling of the challenge where users with (extreme) *facial poses* are recognized.

4. **Towards a real-world end-to-end biometric system.** We design an end-to-end system for thermal-to-visible CFR, which incorporates TFLD and ANYRES as successive operations in the pipeline, towards the ultimate goal of recognition, which can be conducted by any off-the-shelf FR algorithms. The integration of our TFLD as a preprocessing step of ANYRES enables the development of an end-to-end biometric system, from face detection to spectrum translation, that can operate effectively in unconstrained environments. In particular, TFLD is designed to be robust to unconstrained circumstances, while ANYRES offers spectrum translation of any resolution scale, with preserving the identity information, even at a distance or under pose variation. This approach offers a reliable solution in the realm of night-time surveillance and security applications.

Minor contributions

1. **Comprehensive overview of the state of the art.** We provide an extended overview of CFR methods, by firstly formalizing the CFR problem

and secondly describing the general components of a CFR system, viz., face image pre-processing and feature extraction. Finally, we discuss the appropriate spectral bands for FR and elaborate on recent CFR methods, placing emphasis on deep neural networks. In particular we describe techniques that have been proposed to extract and compare heterogeneous features emerging from different spectral bands.

2. **BYDB: A large-scale time synchronized multi-spectral face dataset.** As an industry driven-project, this contribution has to do with the collection of a large-scale multi-spectral face dataset. We present BYDB: Beyond the Visible database, considered as the largest collection at the time of writing of this dissertation, that provides essential, rich and varied data, well-suited to replicate practical real-world scenarios. BYDB is a database comprised of multi-spectral face images captured simultaneously under four electromagnetic spectra, namely the visible, NIR, SWIR and LWIR spectral bands. In this context, the synchronous acquisition ensures the proper alignment of the visible and infrared images, thus enabling a more efficient learning of cross-spectral facial features. The use of this dataset in the development and evaluation of the proposed CFR methods enhances related performance and generalizability under unconstrained environments.

1.5 Thesis outline

The manuscript is composed of seven Chapters, in which Chapter 2, 3, 4, 5 and 6 correspond to details of our contributions. The remaining thesis is organized as follows.

- **Chapter 1: Introduction.** In the first Chapter, we introduce the topic of cross-spectral face recognition. Then we present our goals and discuss related challenges. We finally summarize our contributions.
- **Chapter 2: Literature Review.** In the second Chapter, we present a comprehensive overview of the state of the art in face recognition, beyond the visible spectrum. We discuss appropriate spectral band for CFR and revisit existing CFR methods, as well as associated results and insights.
- **Chapter 3: Multi-spectral Face Dataset.** In the third Chapter, we introduce BYDB, the dataset we have collected, curated and preprocessed.
- **Chapter 4: Face and Landmarks Detector.** In the fourth Chapter, we present an automatic face and landmark detector as a key preprocessing step in CFR.
- **Chapter 5: Identity Preserving Spectrum Translation.** In the fifth Chapter, we adapted generative models to the task of spectrum translation. Particular emphasis is placed on preserving the identity information across spectra, as well as to generate realistic faces.

- **Chapter 6: Towards an end-to-end cross-spectral face recognition system.** In the sixth Chapter, we present a generative model designed to handle simultaneously the dual computer vision tasks of domain translation and super-resolution. Coupled with a prior facial detector algorithm, they provide a reliable solution in the development of an end-to-end biometric system, robust to unconstrained environments.

- **Chapter 7: Conclusions and Perspectives.** In the last Chapter, we summarize the contributions and highlight future lines of research.

1.6 Publications, Patents and Software

The research presented here resulted in five conference papers published in the areas of biometrics and computer vision, as well as three patents. Furthermore, we have submitted one paper to a journal, and developed a demo software to showcase the complete system.

1.6.1 Publications

International conferences

1. David Anghelone, Cunjian Chen, Philippe Faure, Arun Ross, and Antitza Dantcheva. "**Explainable Thermal to Visible Face Recognition Using Latent-Guided Generative Adversarial Network**". In 2021, 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), December 15-18, 2021, Jodhpur, India. [7]

2. Valeriya Strizhkova, Yaohui Wang, David Anghelone, Di Yang, Antitza Dantcheva and Francois Brémond, "**Emotion Editing in Head Reenactment Videos using Latent Space Manipulation**". In 2021, 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), December 15-18, 2021, Jodhpur, India. [124]

3. David Anghelone, Sarah Lannes, Valeriya Strizhkova, Philippe Faure, Cunjian Chen, and Antitza Dantcheva. "**TFLD : Thermal face and landmark detection for unconstrained cross-spectral face recognition**". In 2022, IEEE International Joint Conference on Biometrics (IJCB 2022), October 10-13, 2022, Abu Dhabi, United Arab Emirates. [8]

4. Cunjian Chen, David Anghelone, Philippe Faure, and Antitza Dantcheva. "**Attention-guided generative adversarial network for explainable thermal to visible face recognition**". In 2022, IEEE International Joint Conference on Biometrics (IJCB 2022), October 10-13, 2022, Abu Dhabi, United Arab Emirates. [20]

5. David Anghelone, Sarah Lannes, and Antitza Dantcheva. "**ANYRES : Generating high-resolution visible face images from low-resolution**

thermal-face images". In 2023, IEEE International Conference on Multimedia and Expo (ICME 2023), July 10-14, 2023, Brisbane, Australia. [6]

Submitted Journal paper

1. David Anghelone, Cunjian Chen, Arun Ross, and Antitza Dantcheva. "**Beyond the visible: A survey on cross-spectral face recognition**". arXiv preprint arXiv:2201.04435, 2022. [5]

1.6.2 Patents

1. Anghelone, D. *et al.* "System and Method of Unveiling High-Resolution visible face images from Low-Resolution face images". 2022. Patented EP22306692.9.

2. Anghelone, D. *et al.* "Thermal Face and Landmark detection method". 2022. Patented PCT/EP2022/085480.

3. Anghelone, D. *et al.* "Cross-spectral face recognition training and cross-spectral face recognition method. 2022. Patented PCT/EP2022/085482.

1.6.3 Software

1. Research Days Demo: End-to-End Facial Biometric System operating in the invisible spectra under unconstrained environments.

Chapter 2

Literature Review

Recent advances in convolutional neural networks (CNNs) and generative adversarial networks (GANs) have allowed for remarkable improvement in CFR [21, 148, 47, 34]. In this Chapter, we revisit literature related to the topic of the thesis, and revisit existing CFR-methods, as well as associated results and insights. Unlike other introductory overview studies including [44, 53, 102], we focus on first formalizing the CFR problem in Section 2.1 and discussing the appropriate spectral bands in Section 2.2. Next we describe in Section 2.3 general components of a CFR system, *viz.* from image preprocessing to facial feature extraction for the ultimate task of recognition. Finally, Section 2.4 highlights recent advances based on *deep learning* in addressing CFR, and techniques that have been developed to bridge the modality gap.

2.1 Formalization

Towards enabling comparisons *w.r.t.* existing works, we proceed to formalize the task of CFR. Inspired by previous work [27, 138], we firstly formalize the FR-task and then adapt to the CFR-task as follows.

Let $\mathcal{M}$ be an electromagnetic spectral space modality associated with a marginal distribution $\mathbb{P}$ over a d-dimensional feature space $\mathcal{X} \subset \mathbb{R}^d$ and a label space $\mathcal{Y} \subset \mathbb{N}$.

Given a n-face database $X = \{\mathbf{x}_i\}_{i=1}^n$, where $\mathbf{x}_i \in \mathcal{X}$, and their corresponding n-identities $Y = \{y_j\}_{j=1}^n$ ($y_j \in \mathcal{Y}$), a FR-model is defined as a parametric function $\mathcal{F}_{FR,\Theta}$ described by the random variables X (feature space), Y (label space) and deep learning model parameters, Θ, which extract features from a CNN.

$$\begin{array}{rccc} \mathcal{F}_{FR,\Theta} & : & X \times Y & \rightarrow & [0,1] \\ & & (\mathbf{x}_i, y_j) & \mapsto & \mathbb{P}(Y = y_j | X = \mathbf{x}_i, \Theta), \end{array} \tag{2.1}$$

where, $i, j \in [1, n]$.

Thus, the FR-model aims to deduce parameters, Θ, that increase the probability of correct recognition to 1, assuming that the correct face is in the gallery. More specifically, for all k-index identities $\in [1, n]$

$$\mathcal{F}_{FR,\Theta}(\mathbf{x}_k, y_k) = \mathbb{P}(Y = y_k | X = \mathbf{x}_k, \Theta) = 1. \tag{2.2}$$

However, in the case of CFR, we compare acquired face images in two different modalities (or spectra, or domains): the source $\mathcal{M}^s$ and the target $\mathcal{M}^t$. Each

of them is associated with a marginal distribution $\mathbb{P}$ over the d-dimensional feature space for both source $X^s = \{\mathbf{x}_i^s\}_{i=1}^n \subset \mathcal{X}^s$ and target $X^t = \{\mathbf{x}_i^t\}_{i=1}^n \subset \mathcal{X}^t$, respectively. Note that they share the same set of labels $Y = \{y_j\}_{j=1}^n \subset \mathcal{Y}$.

The CFR-function, $\mathcal{F}_{CFR,\Theta}$, is formalized as follows. Towards finding the deep learning model parameters, Θ, for all k-index identities $\in [1, n]$, we have

$$\mathbb{P}(Y = y_k | X^s = \mathbf{x}_k^s, \Theta) = \mathbb{P}(Y = y_k | X^t = \mathbf{x}_k^t, \Theta) = t, \tag{2.3}$$

where, $t \in [0, 1]$ is a predefined threshold. Algorithms reviewed in this paper seek to provide strategies for determining Θ, formalized in Equation (2.3).

During enrollment, a *gallery face image* is captured and processed and being stored as *reference template*, representing the biometric information of an individual. At the time of authentication, a *probe face image* is captured and processed in the same way as in the enrollment and compared against a reference template of a claimed identity (verification 1:1) or against all stored reference templates (identification 1:N) [69].

Mathematically, both heterogeneous verification and identification are formalized as follows. The user is either *genuine* – the claim is true, or an *impostor* – the claim is false. They are represented by the set $\Omega = \{\omega_{true}, \omega_{false}\}$, respectively.

In *CFR*-verification, an input feature vector $\mathbf{x}_q^t \in \mathcal{X}^t$ stemming from the target modality $\mathcal{M}^t$ and a claimed identity $y_j \in \mathcal{Y}$ are embedded as the pair $(\mathbf{x}_q^t, y_j)$. The comparison algorithm (Section 2.3.2) analyzes the extracted templates $\mathbf{x}_q^t$, and a function $\mathcal{S}$ (Section 2.3.3) computes a *similarity score* (2.8) according to the stored source template $\mathbf{x}_j^s \in X^s$. If the score exceeds a predefined threshold t, it is considered a match. Thus,

$$(\mathbf{x}_q^t, y_j) \in \begin{cases} \omega_{true}, & \text{if } \mathcal{S}(\mathbf{x}_j^s \mathbf{x}_q^t) \geq t \\ \omega_{false}, & \text{otherwise.} \end{cases} \tag{2.4}$$

In *CFR*-identification, an input feature vector $\mathbf{x}_q^t \in \mathcal{X}^t$ derived from the target modality $\mathcal{M}^t$ is provided to the biometric system. The system attempts to search in the gallery for the appropriate identity y_k correlated with $\mathbf{x}_k^s \in X^s$, where $k \in [1, n]$. The associated template is able to provide the highest similarity score (2.8). In addition, the highest score exceeding the predefined threshold t leads to a match. Otherwise the user is classified as unknown $y_{unknown}$. Thus,

$$\mathbf{x}_q^t \in \begin{cases} y_k, & \text{if } \underset{k \in [1,n]}{max} \{\mathcal{S}(\mathbf{x}_k^s, \mathbf{x}_q^t) \geq t\} \\ y_{\text{unknown}}, & \text{otherwise.} \end{cases} \tag{2.5}$$

2.2 CFR in different spectra

CFR involves different spectra, where spectral imaging is marked by specific bands of the electromagnetic spectrum which are responsible of related electromagnetic waves and frequency [156]. These radiations of light consist of a broad range of photon-energy that propagates through electromagnetic waves,

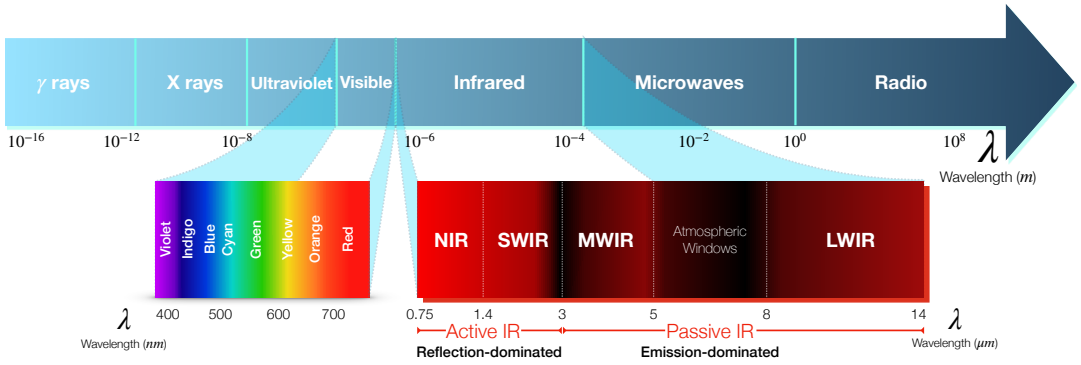


FIGURE 2.1: **Electromagnetic spectrum.** Modalities and their associated wavelengths, highlighting the visible and infrared (IR) radiations.

such as gamma rays, X-rays, ultraviolet radiation, visible light, infrared radiation, microwaves and radio waves.

Although the sun emits its peak power in the visible region, the visible light (which is perceivable by the human eye) including wavelengths from $400nm$ to $750nm$, remains a small portion of the electromagnetic spectrum. Figure 2.1 attempts to classify the electromagnetic radiation described above.

Main interest related to CFR is placed on recognizing faces from the infrared spectrum images. Notably, it offers the capability to acquire images in more challenging environments, including low-light or night-time conditions, or sensing images in atmospherically challenging conditions such as *rain*, *fog*, *mist*, *haze*, and common urban particulates such as *smoke* and *pollution* [10]. We next proceed to introduce the infrared (IR) spectral bands, which we spotlight in Figure 2.1, that have been employed in CFR. Wolff *et al.* [140], Buddharaju *et al.* [15], Kong *et al.* [82], Bhowmik *et al.* [10], and Bourlai and Hornak [13] describe *infrared light* as an invisible, heat-associated energy that can be sensed when radiation or warmth are reflected or emitted from an object. Unlike ultraviolet rays [105], IR-waves penetrate the skin without damage to health. We note that, in principle, IR sensors capture either the face-*reflection* of infrared light or the heat face-*emission* stemming from the subcutaneous superficial blood vessels. In particular, face contains blood capillaries and forms a subcutaneous network. The emission dominated region is therefore referred as the *thermal* infrared spectrum.

IR bands have been defined according to the standard *ISO*-20473:2007 as near-infrared (NIR) between 0.78–3 μm, mid-infrared (MIR) between 3–50 μm and far-infrared (FIR) 50–1000 μm. Deviating from that, the *CIE*-International Commission on Illumination has divided the IR-spectrum into the bands IR-A (0.7 μm – 1.4 μm), IR-B (1.4 μm – 3 μm) and IR-C (3 μm – 1000 μm).

The division of the IR bands is however redefined to follow physical properties from the heat emission of the human body [44] (predicted by the Planck's

law at 305 Kelvin (K), related to different wavelengths), which determine **reflection dominated** and **emission dominated** regions. In particular, a similar scheme has been employed by manufacturers of IR sensors, where specific sensors have been developed in order to respond to electromagnetic radiation constrained by the spectral response. These regions, also denoted as *Passive* and *Active* IR, respectively, are further divided into four sub-spectra: ranging from near infrared (NIR) and shortwave infrared (SWIR) to more challenging bands for FR such as midwave infrared (MWIR) and longwave infrared (LWIR), as ordered in Figure 2.1. In addition, facial images of the same person sensed across the infrared spectrum are depicted in Figure 2.2, where both reflection and emission physiological properties are distinguishable. We note that this dissertation will not cover the *polarimetric thermal imaging*, which involves the measurement of polarization state information of the light contained in the thermal infrared spectrum.

FIGURE 2.2: **Face beyond the visible**. A comparison of a face in visible (VIS) and infrared bands, viz., *NIR*, *SWIR*, *MWIR* and *LWIR*. We note the different physiological properties in both **Active IR** (NIR, SWIR) and **Passive IR** (MWIR, LWIR) bands. Figure credit: Hu *al.* [53].

Further illustrated in Figure 2.3, the emission in the NIR and SWIR sub-bands is nearly zero. Consequently, it requires an external energy to *active* imaging and reveal the scene with an appropriated wavelengths IR light (illuminator). Beyond a wavelength of $\lambda = 0.3$ m, the IR band in the MWIR and LWIR sub-bands is significantly emissive and aims to acquire *passively* heat-sensitive radiation emitted from a human face. Hence, the active IR is characterized by reflective materiel properties, while the passive IR acquires heat emissive energy and is referred as thermal infrared spectrum.

2.3 Components of Face Recognition in *CFR*

Generally, a FR system can be described in multiple stages [70]. The initial stage involves capturing facial images and pre-processing them, which can include tasks like localizing and cropping the faces. In the subsequent stage, features are extracted from the facial images with deep models. These features are then employed in a classifier for the final purpose of identification or verification.

As in classical FR, *face detection* and *alignment* are considered the first and foremost steps of CFR systems. CFR data is preprocessed by aligning and cropping faces to reference templates based on detected facial landmarks.

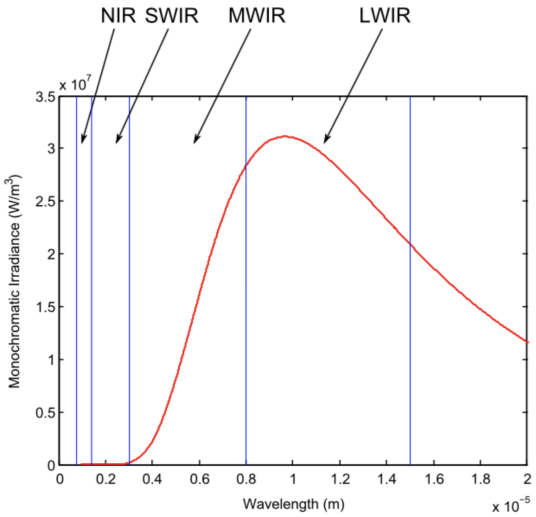


FIGURE 2.3: Curve depicting the Planck's law at 305 K predicting the human body heat emission. The infrared band is divided into four sub-band, namely the NIR, SWIR, MWIR and LWIR respectively, according to the spectral emission. Source [44].

Then, a standard pipeline proceeds in learning/extracting face features by using optimal network architectures and loss functions.

2.3.1 Face Preprocessing

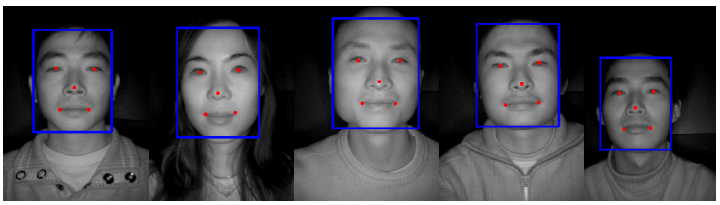
Very limited work has focused on facial landmark detection beyond the visible spectrum [115, 84, 25]. Since NIR and SWIR face images closely resemble visible spectrum faces, landmark detection methods developed for visible face images are adaptable.

Towards demonstrating the possibility of adapting landmark detection methods developed for the visible spectrum, we utilize MTCNN [149] to perform face detection and landmark detection experiments on NIR faces.

Aligning faces and obtaining standardized images is an essential processing step for CFR [44]. In this thesis, we uilize five facial landmarks, namely the geometric center of the eyes, nose and mouth corners, as depicted in Figure 2.4a. Such coordinates are then used to standardize the faces with affine transformations, including translation, scaling, and rotation. The resulted facial images (see Figure 2.4b) are finally canonically aligned, centered and scaled to the same size. We show in Chapter 4 that accurate landmarks used for image face alignment have a strong impact on the recognition scores.

2.3.2 Deep Feature extraction

Once the face is detected on an image and processed into a standardized *aligned* facial image, facial features can be extracted with deep models. In contrast to traditional FR system, CFR allows for two different schemes towards computing domain-invariant features. Firstly, faces from different spectral bands can be directly ingested by a CNN, aiming to learn a shared feature representation



(A) Face and landmark detection

(B) Face aligned

FIGURE 2.4: Face and landmark detection experiments conducted on near-infrared faces sampled from [94]. Face normalization was performed by aligning the faces to the reference pose based on the detected landmarks.

scheme. Secondly, faces can be transformed by GANs into a target domain and features can be extracted from the transformed images.

Convolutional Neural Networks (CNNs)

Formally presented during the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) Computer Vision contest in 2012, *AlexNet* [3] highlighted the *Convolutional Neural Network*, also known as CNNs. However, the premises of these architecture were earlier proposed in 1988 by Fukushima *et al.* [43] and later improved in 1998 by LeCun *et al.*, applying a gradient-based learning algorithm [88].

CNNs are a class of neural networks that are highly efficient in learning effective feature representations for FR [95, 21, 29]. For CFR, CNNs operate on specific pairs of data, *e.g.,* NIR-VIS, formed by different spectral bands. Depending on the number of pairs, CNNs can be designed with a two-branch [118, 48, 49] or a three-branch architecture [96]. Weights from individual branches are often shared. An example of such a design is depicted in Figure 2.5. Visible and NIR images represent respectively input to the network, extracting features. To learn domain-invariant features that are independent of the given spectral bands, a contrastive loss function is applied to enforce the identity constraint for these two extracted feature vectors [118]. For triplet inputs formed by anchor, positive and negative samples, a triplet loss is generally applied to minimize the gap between different spectral bands [96].

There are several prominent CNN-architectures that can be used as backbone for the architecture in Figure 2.5, some of which we enlist in Table 2.1.

Generative Adversarial Networks (GANs)

GANs are deep learning frameworks introduced by Goodfellow *et al.* [45] (2014), which consist of two sub-networks, a *Generative* model G and a *Discriminative*

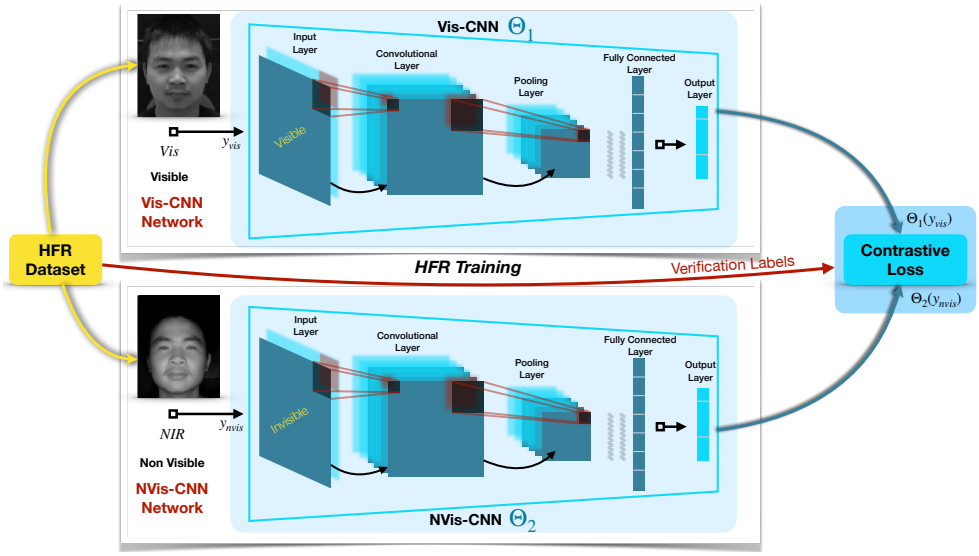


FIGURE 2.5: **CNNs.** Flowchart depicting a two-branch convolutional neural network, learning domain-invariant feature representation.

TABLE 2.1: **Representative architectures** used in CFR for both CNNs and GANs.

Year	Backbone Architecture	Source	Used in
2014	VGGNet	[122]	[91, 11, 27, 78]
	GoogLeNet	[127]	[118]
2015	ResNet	[46]	[17, 30, 24]
2015	U-Net	[119]	[19], [150], [76], [33]
2016	Inception-ResNet	[128]	[58, 27, 54]
2017	DenseNet	[60]	[66, 148]
2018	LightCNN	[141]	[96, 48, 49, 142, 47, 27, 24]

model D trained simultaneously. Together, G and D compete with each other in a minimax game. Specifically, G aims to learn the intrinsic distribution p_G over some target data $x \sim p_{data}$, and associated with a noise prior $z \sim p_z$. G draws sample z for creating synthetic data $G(z)$. In the meantime, D aids the training by taking x or $G(z)$ as input and performing a binary classification as to whether the input is from the real data distribution p_{data} or from the generated data distribution p_G. G attempts to mislead D to not be able to distinguish generated images from real, while D attempts to make that distinction.

The optimization competition of these two models results in the following minimax game

$$\min_G \max_D \mathbb{E}_{x \sim p_{data(x)}}[log(D(x))] + \mathbb{E}_{z \sim p_z}[log(1 - D(G(z)))]. \tag{2.6}$$

Traditional GANs learn a mapping from a random noise input z to an output image y, while conditional GANs learn the mapping from an observed image x and a random noise z in which both G and D are conditioned to an output

image y according to

$$\min_G \max_D \mathbb{E}_{x\sim p_{data(x)}}[log(D(x|y))] + \mathbb{E}_{z\sim p_z}[log(1 - D(G(z|y)))]. \quad (2.7)$$

Many existing techniques have been developed based on conditional GANs to address the CFR problem (see Figure 2.6). The generator takes a thermal image as input, seeking to produce a synthesized visible image as output. The discriminator is trained to distinguish between two pairs of images: the real pair, consisting of an input thermal image and a target visible image, and the synthesized pair, consisting of an input thermal image and a synthesized visible image [150, 33, 21, 34]. As noted in these works, there were notable differences in the use of architectures and loss functions. Some variants of architectures such as multi-scale generator [34] and multi-scale discriminator [148] were used in order to account for scale variance. There also exist variants of loss functions, including attribute loss [34], identity loss [21] and shape loss [137].

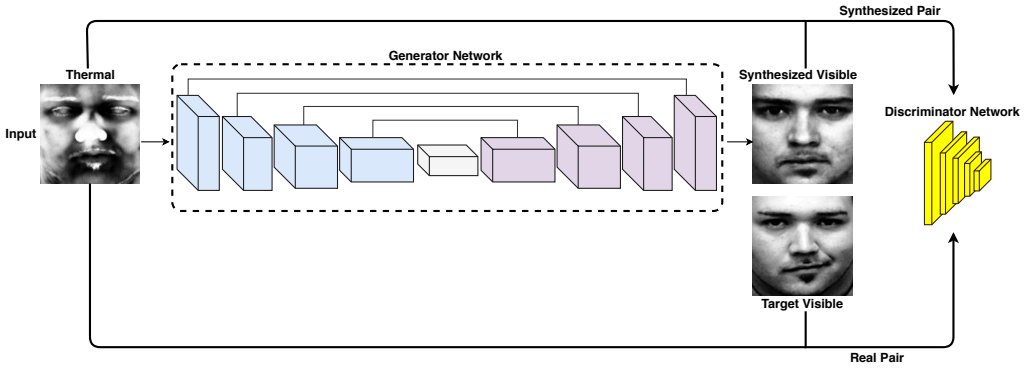


FIGURE 2.6: **GANs.** Flowchart depicting a conditional generative adversarial network for synthesizing visible faces from conditional infrared (thermal) faces.

Loss Functions for *CFR*

CNNs and GANs constitute fundamental deep learning networks. Once such network architectures are designed, suitable loss functions are chosen to minimize prediction errors. Therefore, one active direction in CFR has to do with the development of efficient *loss functions* [101]. Loss functions are pertinent in deep architectures, as they aim to measure how well algorithms are able to perform correct predictions.

In the CFR-context, the loss function is specifically designed to mitigate cross-spectral impact, in order to minimize intra-class variation and maximize inter-class variation [118, 96, 17].

Loss functions used in CFR are frequently adapted from traditional FR. We note that the predominantly used loss function in traditional FR remains to be *softmax*, which is targeted to maximize the probability that the same subject belongs to the target class. Therefore, it encourages the separability of features. *Center loss* [139] seeks to minimize distances between deep features and their corresponding class centers by simultaneously learning a center for individual

classes. However, we note that center loss does not encourage separability of features. Therefore, it is common practice to perform a joint supervision of the softmax loss with center loss. Some other approaches combine euclidean margin-based losses with softmax loss to perform joint supervision. Recently, it has been shown that the features learned by softmax loss have an intrinsic angular distribution [95]. Since then, a flurry of angular-based loss functions have been proposed to impose discriminative constraints on a hypersphere manifold [95, 135, 29]. We summarize loss functions in Table 2.2.

In most traditional FR systems, samples belonging to the same subject are treated without any differentiation. However, in CFR, samples are divided into different categories based on their spectrum information. This prompts the use of loss functions that can accept a pair of inputs [118], triplet of inputs [96], as well as quadruplet of inputs [17]. These loss functions have achieved impressive performance. Admittedly, directly using loss functions from traditional FR for CFR is based on the premise that facial appearance from other spectral bands is close to the visible spectral band. This is applicable in the case of comparing NIR against VIS faces. With increase in wavelength, difference in the facial appearance also starts to increase. Thus, adversarial loss functions used in GAN can be used to perform the synthesis.

2.3.3 Face comparison

Trained CNNs or GANs for CFR produce deep feature representations for a given probe-image, which are then compared with those of the gallery images. Using the same notation introduced in the formalization (Section 2.1), let $X^s = \{\mathbf{x}_i^s\}_{i=1}^n \subset \mathcal{X}^s$ and $X^t = \{\mathbf{x}_i^t\}_{i=1}^n \subset \mathcal{X}^t$ denote the set of samples from the *Source* (visible) and the *Target* (infrared) modalities, respectively. The corresponding shared set of labels are denoted by $Y = \{y_j\}_{j=1}^n \subset \mathcal{Y}$.

Suppose Θ denotes the deep process of extracting d features from a CNN, then the similarity score between two templates can be computed as a function $\mathcal{S}$ referred to as the *measure of similarity* ranging between 0 and 1, where 1 represents a high similarity. As an example, the *cosine distance* is widely used to calculate such similarity, for all k-index identities $\in [1, n]$:

$$\begin{aligned} \mathcal{S}(\mathbf{x}_k^s, \mathbf{x}_k^t) = \cos(\mathbf{s}, \mathbf{t}) &= \frac{< \mathbf{s}, \mathbf{t} >}{\|\mathbf{s}\|\|\mathbf{t}\|} \\ &= \frac{\sum_{j=1}^d s_j t_j}{\sqrt{\sum_{j=1}^d (s_j)^2}\sqrt{\sum_{j=1}^d (t_j)^2}}, \end{aligned} \tag{2.8}$$

where, $\mathbf{s} = \Theta(\mathbf{x}_k^s)$ and $\mathbf{t} = \Theta(\mathbf{x}_k^t)$. Other Euclidean distance based measures such as L_2 and L_1 can also serve as similarity metrics.

We proceed to use different Θ models as state-of-the-art facial features extractors. Among these, *SphereFace* [95], *AM-Softmax* [21], *MobileFaceNet* [21]

TABLE 2.2: **Loss Functions.** A comprehensive list of loss functions $\mathcal{L}$ from traditional FR to CFR. Here, N is the number of samples, T is the number of samples in a mini-batch, W is the weight matrix, b is the bias term, x_i and y_i are the i^{th} training sample and the according class label, respectively. $\theta_{y_i} \in \left[\frac{k\pi}{m}, \frac{(k+1)\pi}{m}\right]$ with $k \in [0, m-1]$ and $m \geq 1$.

Task	Name	Notation	Reference	Function $\mathcal{L}$
FR	Softmax (Cross-Entropy)	$\mathcal{L}_s$	[125]	$-\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}\right)$
	Center Loss	$\mathcal{L}_{ce}$	[139]	$\frac{1}{2}\sum_{i=1}^{N}\|x_i - c_{y_i}\|_2^2$
	Marginal Loss	$\mathcal{L}_{ma}$	[28]	$\frac{1}{m^2 - m}\sum_{i,j,i\neq j}^{m}\left(\xi - y_{ij}\left(\theta - \|\frac{x_i}{\|x_i\|} - \frac{x_j}{\|x_j\|}\|_2^2\right)\right)$
	Angular Softmax Loss (SphereFace)	$\mathcal{L}_{as}$	[95]	$-\frac{1}{N}\sum_{i=1}^{N}\frac{e^{\|x_i\|\cos(m\theta_{y_i})}}{e^{\|x_i\|\cos(m\theta_{y_i})} + \sum_{j=1, j\neq y_i}^{n} e^{\|x_i\|\cos(\theta_j)}}$
	Large Margin Cosine Loss (CosFace)	$\mathcal{L}_{co}$	[135]	$-\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{s(\cos(\theta_{y_i}) - m)}}{e^{s(\cos(\theta_{y_i}) - m)} + \sum_{j=1, j\neq y_i}^{n} e^{s\,\cos(\theta_j)}}\right)$
	Additive Angular Margin Loss (ArcFace)	$\mathcal{L}_{aa}$	[29]	$-\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{s\,\cos(\theta_{y_i} + m)}}{e^{s\,\cos(\theta_{y_i} + m)} + \sum_{j=1, j\neq y_i} e^{s\,\cos(\theta_j)}}\right)$
CFR	Contrastive Loss	$\mathcal{L}_{ct}$	[118] - (2.9)	$\begin{cases}\frac{1}{2}\|f(x_i) - f(x_j)\|^2, & \text{if } y_i = y_j \\ \frac{1}{2}max(0, m - \|f(x_i) - f(x_j)\|)^2, & \text{else.}\end{cases}$
	Triplet Loss	$\mathcal{L}_{tp}$	[96] - (2.10)	$\sum_{i=1}^{N}(\|f(x_i^a) - f(x_i^p)\| - \|f(x_i^a) - f(x_i^n)\| + m)$
	Tetrad Margin Loss	$\mathcal{L}_{TML}$	[17] - (2.12)	$\sum_{i\in\{N,V\}}^{T}\left(\frac{z_j^N \cdot z_j^V}{\|z_j^N\| \cdot \|z_j^V\|} - \frac{z_j^N \cdot z_l^V}{\|z_j^N\| \cdot \|z_l^V\|} + m_1\right) + \sum_{i\in\{N,V\}}^{T}\left(\frac{z_j^N \cdot z_j^V}{\|z_j^N\| \cdot \|z_j^V\|} - \frac{z_j^V \cdot z_l^V}{\|z_j^V\| \cdot \|z_l^V\|} + m_2\right)$

and *ArcFace* [29] provide deep characteristic from a visible face and the thermal counterpart face of the Figure 2.7, which are then compared with the cosine similarity $\mathcal{S}$. Table 2.3 report the performances achieved by each models.

FIGURE 2.7: Example images of a subject captured in the thermal infrared (LWIR) sub-band and the visible spectrum. Source from the ARL-MMFD dataset [52].

TABLE 2.3: Computation of cosine similarity $\mathcal{S}$ scores using state-of-the-art face matchers Θ from a visible spectrum image and its thermal spectrum counterpart (facial images from Figure 2.7). The scores are normalized to [0,1].

Θ	Visible vs. Visible	Thermal vs. Visible
SphereFace [95]	1.0	0.499
AM-Softmax [21]	1.0	0.543
MobileFaceNet [21]	1.0	0.505
ArcFace [29]	1.0	0.641

From Table 2.3, every Θ models report a perfect cosine similarity $\mathcal{S}$ scores when facial images of the same person are compared within the visible spectrum, *i.e.,* Visible vs. Visible. However, similarity $\mathcal{S}$ dramatically decreases when we attempt to compare facial images sensed in different spectra. Here, ArcFace-based FR matcher [29] report the higher score of similarity $\mathcal{S}$, thereby motivated our choice to use ArcFace model as facial feature extractor along with the cosine similarity as a metrics throughout the thesis for assessing our CFR algorithms.

Considered as the state-of-the-art during the time of writing this dissertation, ArcFace [29] outperformed several FR benchmarks such as LFW and YTF datasets, *w.r.t.* related accuracy of 99.83% and 98.02% [136], respectively. It is known for its high accuracy and robustness, particularly in dealing with large variations in poses, illumination, and occlusions. ArcFace relies on deep neural network, especially on the ResNet-based architecture to extract facial features for the purpose of FR, including *identification* and *verification*. The loss function has also been updated, from a traditional Softmax loss to Additive Angular

Margin loss version (see Table 2.2) which enforces a larger margin between different classes, resulting in improved discrimination power. Therefore, we utilize a pre-trained ArcFace recognition network to extract facial feature embedding and measure the cosine similarity. The network was trained on normalized face images of size 112×112 from MS-Celeb-1M dataset, with additive angular margin loss. ResNet-50 was used as the embedding network and the final embedded feature size was set to 512.

We note that in the context of this CIFRE-thesis, Thales uses the in-house FR comparison (matcher), namely the Facial Recognition Platform (FRP) [38].

2.3.4 Biometric evaluation metrics

In order to evaluate the performance of CFR systems, several metrics are commonly used in biometrics [69, 70] in which we proceed to enlist.

The first metric is the *False Acceptance Rate* (FAR), that indicates the probability of an imposter being wrongly accepted as a genuine user. This means that the system has made an error in identifying an individual as someone else. Minimizing FAR is critical as it is a measure of the security of the system and reducing it is important to avoid security issues.

The second metric commonly used in biometrics is the *False Rejection Rate* (FRR), that refers to the probability that a genuine user is incorrectly rejected as an imposter. This means that the system has failed to correctly identify an individual who is already enrolled in the system. FRR is a critical metric because it measures the accuracy of the system, and it is essential to minimize it to prevent user frustration and inconvenience of users.

Combining the FAR and FRR, the *Receiver Operating Characteristic* (ROC) curve is essential in FR, as it displays the trade-off between FAR and FRR. It is represented graphically and helps to evaluate the system's performance at different operating points. Notably, the *Area Under Curve* (AUC) of the ROC curve is responsible to evaluate the overall performance of the system. The point of the ROC curve depicting the intersection of the FAR and FRR expresses the *Equal Error Rate* (EER). Note that AUC and EER are metrics especially useful in comparing the performance of different FR systems.

Overall, the choice of metrics usually depends on the specific application of FR, as well as the performance requirements and constraints of the system. Evaluation of models implemented in this thesis will be assessed with these presented metrics.

2.4 *CFR* algorithms of different spectra

Face recognition operating on imagery beyond the visible spectrum can be categorized as

- Cross-spectrum *featured-based* algorithms.

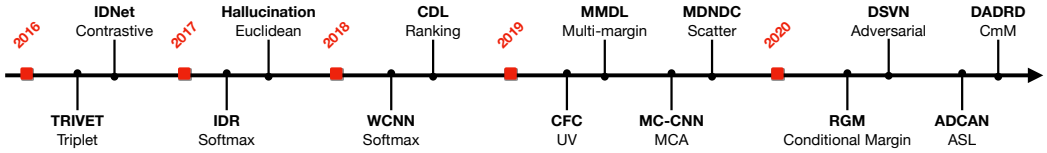


FIGURE 2.8: **Timeline of algorithm development.** NIR-to-VIS CFR with their corresponding loss functions.

- Cross-spectrum *images synthesis* algorithms.

The former involve comparison of infrared probe against a gallery of visible face images, within a common feature subspace. However, after the success of image synthesis based on GANs, cross-spectrum images synthesis algorithms have attempted to synthesize a "pseudo-visible" image from an infrared image. Given the synthesized face image, both academic and commercial-off-the-shelf (COTS) FR-systems trained on visible spectrum, can be utilized for comparison.

We proceed to present CFR algorithms, according to the spectral bands in which the infrared probe images were acquired.

2.4.1 Reflective IR-to-visible *FR*

NIR-VIS

The appearance difference between NIR and VIS face images is less pronounced compared to other spectral bands and, hence, shared feature representations have often been proposed [96, 118, 48, 49, 142]. CNNs are suitable for NIR-to-VIS CFR scenario, since they seek to automatically extract representative face features. Nevertheless, a challenge remains, which has to do with the illumination-variation between the two modalities. We proceed to discuss a number of recent work, constituting prominent state-of-the-art approaches in matching NIR against VIS face images. These algorithms along with their individual loss functions are summarized in Figure 2.8. Further, we also list in Table 2.4 the NIR-to-VIS face comparison performances on public benchmark datasets.

TRIVET. Liu *et al.* [96] proposed a CNN with ordinal measures (o-CNN), pre-trained on visible images from the large-scale CASIA WebFace dataset [144] and hence is able to extract general face features, which has been fine-tuned on pairs of NIR-VIS face images, in order to learn a domain-invariant deep representation. To cope with limited image pairs, two types of NIR-VIS triplet loss functions were used, reducing intra-class variations by iteratively setting NIR and VIS images as anchors, such that the network focuses on the identity distinction instead of the spectrum classification.

IDNet. Reale *et al.* [118] utilized GoogLeNet networks with small convolutional filters, pre-trained on the visible CASIA WebFace dataset. For NIR-to-VIS CFR scenario, two identical networks were initialized based on the pre-trained network by excluding fully connected and softmax layers. The outputs from these two networks, *i.e.,* VisNet and NIRNet, were concatenated to create

TABLE 2.4: Rank-1 Accuracy and Verification Rate on different datasets for the **NIR-VIS** face comparison.

Authors	**Method**	**Loss**	**Dataset**	**Performance**		
				Rank-1	FAR 1%	FAR 0.1%
Liu *et al.* [96]	TRIVET	Triplet	CASIA NIR-VIS 2.0 [94]	95.74	98.1	91.03
			Oulu-CASIA NIR-VIS [22]	92.2	67.9	33.6
			BUAA-VisNir face [59]	93.9	93.0	80.9
Reale *et al.* [118]	IDNet	Contrastive	CASIA NIR-VIS 2.0 [94]	87.1	-	74.5
			CASIA HFB [93]	97.58	96.9	85.0
He *et al.* [48]	IDR	Softmax	CASIA NIR-VIS 2.0 [94]	97.3	98.9	95.7
			Oulu-CASIA NIR-VIS [22]	94.3	73.4	46.2
			BUAA-VisNir face [59]	94.3	93.4	84.7
Lezama *et al.* [91]	Hallucination	Euclidean	CASIA NIR-VIS 2.0 [94]	96.41	-	-
He *et al.* [49]	WCNN	Softmax	CASIA NIR-VIS 2.0 [94]	98.7	99.5	98.4
			Oulu-CASIA NIR-VIS [22]	98.0	81.5	54.6
			BUAA-VisNir face [59]	97.4	96.0	91.9
Wu *et al.* [142]	CDL	Ranking	CASIA NIR-VIS 2.0 [94]	98.6	-	98.32
			Oulu-CASIA NIR-VIS [22]	94.3	81.6	53.9
			BUAA-VisNir face [59]	96.9	95.9	90.1
He *et al.* [47]	CFC (GAN)	UV	CASIA NIR-VIS 2.0 [94]	98.6	99.2	97.3
			Oulu-CASIA NIR-VIS [22]	99.9	98.1	90.7
			BUAA-VisNir face [59]	99.7	98.7	97.8
Cao *et al.* [17]	MMDL	Multi-margin	CASIA NIR-VIS 2.0 [94]	99.9	-	99.4
			Oulu-CASIA NIR-VIS [22]	100.0	-	97.2
Deng *et al.* [30]	MC-CNN	MCA	CASIA NIR-VIS 2.0 [94]	99.22	-	99.27
Hu *et al.* [58]	MDNDC	Scatter	CASIA NIR-VIS 2.0 [94]	98.9	99.6	97.6
			Oulu-CASIA NIR-VIS [22]	99.8	88.1	65.3
Cho *et al.* [24]	RGM	Conditional-margin	CASIA NIR-VIS 2.0 [94]	99.3	99.51	99.02
			BUAA-VisNir face [59]	99.67	99.22	-
Hu *et al.* [56]	DSVN	Adversarial	CASIA NIR-VIS 2.0 [94]	99.0	99.7	98.6
			Oulu-CASIA NIR-VIS [22]	100.0	99.3	95.5
Iranmanesh *et al.* [66]	CpGAN	Contrastive+Coupled +GAN+Eucl.+Perceptual	CASIA NIR-VIS 2.0 [94]	96.63	-	87.05
			CASIA HFB [93]	99.64	98.4	89.7
Hu *et al.* [54]	ADCANs	ASL	CASIA NIR-VIS 2.0 [94]	99.1	99.6	98.5
			Oulu-CASIA NIR-VIS [22]	99.8	93.2	78.9
			BUAA-VisNir face [59]	99.8	99.7	98.4
Hu *et al.* [57]	DADRD	CmM	CASIA NIR-VIS 2.0 [94]	99.1	99.6	98.6
			Oulu-CASIA NIR-VIS [22]	100.0	98.5	92.9
			BUAA-VisNir face [59]	99.9	99.9	99.8

a siamese network with contrastive loss. Such a coupled deep network design was able to map NIR and VIS faces into a spectrum-independent feature space.

IDR. He *et al.* [48] presented a CNN-based approach targeted to map both NIR and VIS images into a common subspace by separating the feature space into a shared layer: NIR- and VIS-layer. While the shared layer encoded the modality-invariant identity information, the NIR- and VIS-layers encoded the modality-variant spectrum information, respectively. The NIR and VIS images were respectively inputs to the two CNN channels with shared parameters. Compared to similar CNN pipelines [96, 118], the proposed method jointly learned identity and spectrum information, leading to substantial performance gains on the CASIA NIR-VIS 2.0 dataset.

Hallucination. Lezama *et al.* [91] proposed an approach, which adapts a pre-trained VIS-CNN model towards generating discriminative features for both VIS and NIR face images, without retraining the network. Their approach consisted of two core components, cross-spectral hallucination and low-rank embedding, aiming at modifying both inputs and outputs of the CNN for CFR. Firstly, cross-spectral hallucination was used to transform the NIR image into the VIS spectrum by learning correspondences between NIR and VIS patches. Next, a low-rank embedding was used to restore a low-rank structure that simultaneously minimized the intra-class variations, while maximizing the inter-class variations. Three pre-trained VIS-CNN models, including *VGG-S*, *VGG-face* and *COTS*, were evaluated on the CASIA NIR-VIS 2.0 dataset.

WCNN. He *et al.*[49] proposed a novel method entitled *Wasserstein Convolutional Neural Network* (WCNN) streamlined to learn invariant features between NIR and VIS face images. Similar to their previous work [48], the proposed WCNN divided the high-level layer of the shared CNN into two orthogonal subspaces that contain modality-invariant identity features and modality-variant spectrum features. In addition, Wasserstein distance was used to measure the difference between heterogeneous feature distributions to reduce the modality gap. We note that their work represented the first attempt to formulate a probability-based distribution learning for VIS-to-NIR face comparison. WCNN outperformed the work of He *et al.* [48], improving the performance on the CASIA NIR-VIS 2.0 dataset.

CDL. Wu *et al.* [142] designed a *Coupled Deep Learning* (CDL) approach, identifying a shared feature space, in order to reduce modality differences in heterogeneous face comparison. The loss function of CDL consisted of two different components: (a) relevance constraints that imposed a trace norm to the softmax loss and a block-diagonal prior to the fully connected layer; and (b) cross-modal ranking that employed triplet ranking regularization to enlarge the size of training data. A semi-hard triplet selection method was also adopted to further improve the NIR-to-VIS comparison performance. Both relevance and ranking losses were jointly optimized.

CFC. He *et al.* [47] proposed a GAN, termed *Cross-spectral Face Completion* (CFC), to synthesize VIS images from NIR images. The heterogeneous face synthesis process was divided into two complementary components: (a) a texture inpainting component that aims to recover a VIS facial texture map from a given NIR image; and (b) a pose correction component that aims to map any

pose in NIR images to the frontal pose in VIS images. A warping procedure was used to fuse these two components in the adversarial network, which was regularized by the combinations of UV loss, adversarial loss and L1 reconstruction loss. In summary, the main contributions of this work had to do with creating a novel GAN-based end-to-end deep framework for cross spectral face synthesis without assembling multiple image patches. It contained encoder-decoder structured generators and two novel discriminators to fully consider variations of NIR and VIS images. This was the first time that unsupervised heterogeneous face synthesis problem was simplified to a one-to-one image translation problem. The decomposition of texture inpainting and pose correction enables the generation of realistic identity preserving VIS face images possible.

MMDL. Cao *et al.* [17] combined heterogeneous representation network and decorrelation representation learning in order to design a *Multi-Margin based Decorrelation Learning* (MMDL) framework to extract decorrelated features for both VIS and NIR images. First, the heterogeneous representation network was initialized from a VIS face model and fine-tuned on both VIS and NIR face images. Second, a decorrelation layer was appended to shared feature layer of the representation network to extract decorrelated features for both VIS and NIR images.

MC-CNN. Deng *et al.* [30] proposed a *Mutual Component Convolutional Neural Network* (MC-CNN) to extract modality-invariant features. The novelty of MC-CNN was the incorporation of a generative module, *i.e.,* the mutual component analys (MCA), into the CNN by replacing the fully connected layer with MCA. VIS-NIR image pairs from the same subjects were first input to a shared CNN to extract common feature vectors. Subsequently, these feature vectors were fed into the MCA layer. Finally, MCA loss was included as an additional regularization to the softmax loss.

MDNDC. Hu *et al.* [58] proposed *Multiple Deep Networks with scatter loss and Diversity Combination* (MDNDC). They used three ResNet-v1 networks; one being dedicated to feature extraction, in order to build the architecture of Multiple Deep Networks (MDN), followed by the two other networks in parallel. Jointly with the MDN, a Scatter loss (SL) [55] was used as a loss function, which attempts to maximize distance between the classes and minimize distance within the class to learn highly discriminative features for the CFR task. In addition to the MDN, a joint decision strategy named Diversity Combination (DN) was introduced to auto-adjust weights of all three deep networks of the MDN and make a joint classification decision.

RGM. Cho *et al.* [24] were interested in Relational Deep Feature Learning for CFR. Therefore, to bridge the domain gap between visible and invisible spectrum, they proposed a *Relational Graph-structured Module* (RGM) focused on facial relational information. Their graph RGM performed relational modeling from node vectors that represent facial components such as lips, nose and chin. They also performed recalibration by considering global node correlation via *Node Attention Unit* (NAU) to focus on the more informative nodes arising from the relation-based propagation. Furthermore, they suggested a novel conditional-margin loss function *C-Softmax* for efficient projection learning on the common latent space of the embedding vector that adaptively uses the

inter-class margin.

DSVN. Hu *et al.* [56] proposed *Disentangled Spectrum Variations Networks* (DSVN) to disentangle spectrum variations between VIS and NIR domains and to separate the modality-invariant identity information from modality-variant spectrum information. DSVN is comprised of two major components: spectrum-adversarial discriminative feature learning (SaDFL) and step-wise spectrum orthogonal decomposition (SSOD). The SaDFL was further divided into identity-discriminative subnetwork (IDNet) and auxiliary spectrum adversarial subnetwork (ASANet). The IDNet was used to extract identity discriminative features. The ASANet, on the other hand, was used to eliminate modality-variant spectrum information. Both IDNet and ASANet were designed to extract domain-invariant feature representations via adversarial learning.

ADCANs. Hu *et al.* [54] proposed an effective *Adversarial Disentanglement spectrum variations and Cross-modality Attention Networks* (ADCANs) for the VIS-to-NIR CFR scenario. To reduce the gap of cross-modal images and solve the NIR-VIS CFR problem, the authors set-up three key components which are able to learn identity-related and modality-unrelated features. Firstly, they proposed a new objective loss termed *Advanced Scatter Loss* (ASL), aiming at capturing within- and between-class information of the data and embedding them in the network for more effective training, focusing on categories with small inter-class distance and increasing the distance between them. Then, a *Modality-adversarial Feature Learning* (MaFL) including an Identity-Discriminative Feature Learning Network (IDFLN) and a Modality-Adversarial Disentanglement Network (MADN) were incorporated to improve the feature representation of the identity-discriminative component as well as to highlight the spectrum variations through adversarial learning. Finally, a *Cross-modality Attention Block* (CmAB) was introduced, in order to guide the network in selecting pertinent features and suppress noise information. CmAB sequentially applies spatial and channel attention modules to both, IDFLN and MADN, in order to increase the representation ability between them.

DADRD. Hu *et al.* [57] proposed the *Dual Adversarial Disentanglement and deep Representation Decorrelation* (DADRD) approach to reduce the gap between NIR-VIS modalities, while enhancing the learning of identity-related features. At the same time, DADRD sought to effectively disentangles the additional residual-related features (*i.e.,* poses, illuminations and expressions (PIE)) rather than only extracting modality-related (*i.e.,* NIR and VIS) and identity-related features. In contrast with their prior work, MDNDC [58] and ADCAN [54], the DADRD method combined three key components. Firstly, they proposed a new objective loss termed *Cross-modal Margin* (CmM), and attempted to enhance the learning identity-related features, which captures within- and between-class information of the data (*i.e.,* occlusion, pose, distance, lighting and expressions), while also reducing the modality gap by using a center-variation item. Then, they introduced a *Mixed Facial Representation* (MFR) which is divided into three layers: “(I) Identity-related layer", “(M) Modality-related layer" and “(R) Residual-related layer". A *Dual Adversarial Disentanglement Variation* (DADV) was then designed to reduce the intra-class

variation with the help of adversarial learning, including both *Adversarial Disentangled Modality Variations* (ADMV) and *Adversarial Disentangled Residual Variations* (ADRV). These adversarial mechanisms disentangle the spectrum variation (in terms of data heterogeneity) and eliminate residual variations such as PIE, respectively. Finally, a *Deep Representation Decorrelation* (DRD) was embedded to the three MFR layers of DADV for the purpose of making them unrelated to each other and enhance feature representations. The DADRD method, combining CmM, DADV and DRD into a unified framework, is depicted in Figure 2.9.

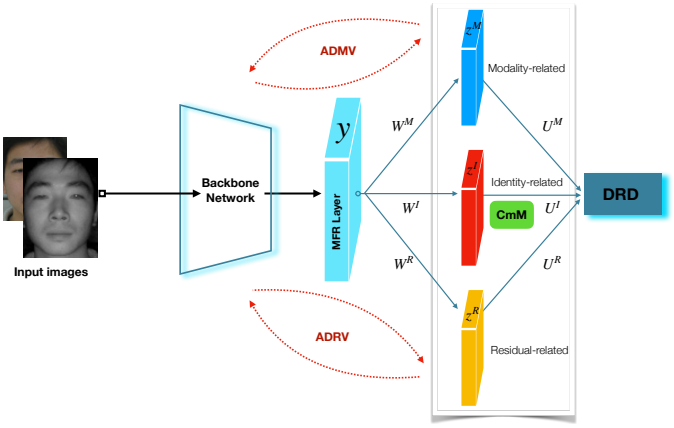


FIGURE 2.9: **Overview of the DADRD [57] model:** learn identity discriminative features and disentangle within-class variations. DADRD combines three key components, including *CmM* loss, *DADV* consisting of both *ADMV* and *ADRV* and *DRD* embedded into the three previous *MFR* layers.

DVG-Face. Fu *et al.* [42] proposed a novel Dual Variational Generation (DVG-Face) framework to learn the joint distribution of paired heterogeneous images. Identity information from large-scale visible face data was also utilized to tackle the problem of identity diversity sampling induced by the small-scale paired heterogeneous data. To ensure identity consistency, a pairwise identity preserving loss was applied to generated paired heterogeneous images. A contrastive learning scheme was employed to train the heterogeneous face matching with positive pairs from generated paired heterogeneous images and negative pairs from different identity samplings. The proposed DVG-Face was an extension from the authors' previous work DVG [41] that greatly improved the identity diversity of generated face images.

Summary of NIR-to-VIS Face Comparison. Comparing face images in NIR against VIS can be coarsely classified into (a) cross-spectrum featured-based methods [96, 118, 48, 49, 142, 17, 30, 24, 56] and (b) cross-spectrum image-synthesis methods [91, 47]. Existing methods (a) predominantly seek to learn a domain-invariant deep representation between NIR and VIS samples, which is motivated by the presence of some similar textures observed in NIR as well as VIS-domains. Such approaches can incorporate two different ways to construct the inputs of the networks. Firstly, NIR and VIS images pertaining to same subjects are related in positive pairs [118, 48, 49, 17, 30, 56], constituting

two identical subnetworks with shared weights. A *contrastive loss* in this context minimizes the intra-class distance

$$\mathcal{L}_{ct} = \begin{cases} \frac{1}{2}\|f(x_i) - f(x_j)\|^2, & \text{if } y_i = y_j \\ \frac{1}{2}max(0, m - \|f(x_i) - f(x_j)\|)^2, & \text{else.} \end{cases} \tag{2.9}$$

Secondly, a triplet can be formed, where a NIR image is set as an anchor, a VIS image of the same subject as a positive sample, and an additional VIS image pertaining to a different subject as a negative sample. Alternatively, a triplet can be formed by setting a VIS image as an anchor and using NIR images from the same and different subjects as positive and negative samples, respectively [96, 142]. This leads to three identical subnetworks with shared weights, where the *triplet loss* is defined as

$$\mathcal{L}_{tp} = \sum_{i}^{N} \left(\|f(x_i^a) - f(x_i^p)\| - \|f(x_i^a) - f(x_i^n)\| + m\right). \tag{2.10}$$

Here, x_i^a is an anchor image, x_i^p is a positive sample and x_i^n is a negative sample. The goal of triplet loss is to 'push' the negative sample x_i^n away from the anchor x_i^a by a margin m compared to the positive sample:

$$\|x_i^a - x_i^p\|^2 + m \leq \|x_i^a - x_i^n\|. \tag{2.11}$$

The triplet loss attempts to minimize the intra-class distance, while maximizing the inter-class distance. Unlike triplet, where a group of three samples is selected, the *Tetrad Margin Loss* (TML) [17] uses a group of four samples to construct tetrad tuples. The designed tetrad sample selection strategy was to choose four heterogeneous decorrelation representations, viz., $\{z_j^N, z_j^V, z_k^N, z_l^V\}$. Herein, $\{z_j^N, z_j^V\}$ are samples from the same identity, z_k^N denotes the closest NIR sample to z_j^V from another identity, z_l^V represents the closest VIS sample to z_j^N from another identity. The tetrad margin loss is defined as:

$$\begin{aligned} \mathcal{L}_{TML}(z_j^N, z_j^V, z_l^V) = & \sum_{i\in\{N,V\}}^{T} \left(\frac{z_j^N \cdot z_j^V}{\|z_j^N\| \cdot \|z_j^V\|} - \frac{z_j^N \cdot z_l^V}{\|z_j^N\| \cdot \|z_l^V\|} + m_1 \right) \\ & + \sum_{i\in\{N,V\}}^{T} \left(\frac{z_j^N \cdot z_j^V}{\|z_j^N\| \cdot \|z_j^V\|} - \frac{z_j^V \cdot z_l^V}{\|z_j^V\| \cdot \|z_l^V\|} + m_2 \right). \end{aligned} \tag{2.12}$$

Subsequently, $\mathcal{L}_{TML}(z_j^N, z_j^V, z_k^N, z_l^V)$ can be defined as

$$\begin{aligned} \mathcal{L}_{TML}(z_j^N, z_j^V, z_k^N, z_l^V) = & \mathcal{L}_{TML}(z_j^N, z_j^V, z_l^V) \\ & + \mathcal{L}_{TML}(z_j^V, z_j^N, z_k^N). \end{aligned} \tag{2.13}$$

T represents the number of samples in a mini-batch, while m_1 and m_2 are the two margins. The tetrad margin loss can be regarded as the combination of two triplet losses [96, 142]. Intuitively, $\mathcal{L}_{TML}(z_j^N, z_j^V, z_l^V)$ is very similar to assigning a NIR image z_j^N as an anchor, a VIS image z_j^V of the same subject as a positive sample, and another VIS image z_l^V of a different subject as a negative sample. The difference lies in the use of cosine distance other than the euclidean

distance to compare the similarity. Both triplet loss and tetrad margin loss use the hard-sample mining strategy.

In addition to the contrastive loss, triplet loss and tetrad margin loss, other loss functions have notably been proposed in the NIR-to-VIS literature. In [48], a variant of *softmax loss* was proposed in order to learn modality invariant subspace. The softmax loss function is defined as:

$$\begin{aligned} \mathcal{L} &= \sum_{i \in \{N,V\}} \text{softmax}(F_i, c, \Theta, W, P_i) \\ &= -\sum_{i \in \{N,V\}} \left(\sum_{j=1}^{N} \log(\hat{p}_{ij}) \ \mathbb{1}_{\{y_{ij}=c\}} \right), \\ s.t. \quad & P_i^T W = 0, \quad i \in \{N, V\}, \end{aligned} \tag{2.14}$$

where, F_i is the fully connected layer of WCNN, c is the class label and Θ is the set of WCNN parameters. W is used to denote the weight matrix of the modality-invariant features (*i.e.,* shared features across spectral domain) and P_i denotes the weight matrix of the spectrum-specific features (*i.e.,* features which are spectral dependent). $\mathbb{1}$ is an indicator function and p_{ij} denotes the predicted class probability. $P_i^T W$ corresponds to an orthogonal constraint imposed to make the features uncorrelated.

SWIR-VIS

To date, SWIR-to-visible CFR scenario has received limited attention. This is reflected in the scarcity of deep models and in the lack of publicly available SWIR face datasets (see Chapter 3). We proceed to summarize existing prominent handcrafted methods [13, 12].

pre-CNN. Research of Bourlai *et al.* [14] firstly investigated face verification related to the SWIR band. The authors introduced a *geometric* and *photometric normalization* (PN) scheme in order to compensate for the variable environment, coupled with *contrast limited adaptive histogram equalization* (CLAHE). Based on that, three FR methods (commercial and academic) were tested using the *West Virginia University* (WVU) Multi-spectral database.

Kalka *et al.* [75] extended the previous work *w.r.t.* pre-processing [14] by adding *single scale retinex* (SSR) and also explored the cross-photometric score level fusion rule. For the experiment, three datasets were considered including different environmental setup (fully-controlled/semi-controlled/uncontrolled). Their finding provided insights such as that pre-processing with PN improved FR performance, obtaining highest results up to Rank-1 accuracy of 100%, when images were acquired under fully controlled conditions.

Bourlai *et al.* [12] pursued the intra-spectral and cross-spectral FR specially focused on (NIR/SWIR/MWIR)-imaging at various standoff distances and different environmental conditions. They merged previous work [14, 75] in order to demonstrate that jointly, the use of independent or combined PN, CLAHE, SSR and cross-photometric score level fusion rule were able to reach better results under different environmental setups.

TABLE 2.5: Rank-1 Accuracy and Verification Rate on different dataset for the **MWIR-VIS** face comparison.

Authors	Methods	Dataset	Performance %
Sarfraz *et al.* [120]	DPM	NVESD [16]	98.66 @Rank-1
Zhang *et al.* [150]	TV-GAN	IRIS [132]	19.90 @Rank-1
Chen *et al.* [21]	SG-GAN	PCSO [81]	92.16 @AUC - 15.01 @EER
Iranmanesh *et al.* [66]	CpGAN	NVESD [16]	96.10 @Rank-1
		WSRI [26]	97.80 @Rank-1
Peri *et al.* [112]	Pix2Pix-ATC	MILAB-VTF(B) [112]	59.3 @AUC - 43.4 @EER
	CycleGAN-ATC	MILAB-VTF(B) [112]	54.9 @AUC - 46.4 @EER
	CUT-ATC	MILAB-VTF(B) [112]	68.8 @AUC - 36.3 @EER

CMLD. Cao *et al.* [18] proposed a local operator called *Composite Multilobe Descriptors* (CMLD) with the aim to extract facial feature through all IR-spectral bands. This operator combined Gaussian function, Local binary patterns *(LPB)*, Weber local descriptor and histogram of oriented gradients *(HOG)*. The experimental results showed that for the SWIR-to-visible scenario, CMLD performed on a private dataset containing SWIR images (at various standoff distances), a Rank-1 accuracy of 78.7% and a verification rate of 99.5% at FAR 10%. Authors also used CMLD to conduct a study on cross-spectral *partial* FR in order to understand the face area which contained useful information for CFR. They found that the eye region is most informative for FR beyond the visible.

VGGFace. Bihn *et al.* [11] examined the ability of using pretrained VGG-Face network on visible images to extract features on SWIR images. Hypothesized was that the VGG-Face network would perform well on face images taken from both, VIS and SWIR wavelengths since their facial appearance difference is not as significant as these images captured at MWIR and LWIR wavelengths. SWIR wavelengths at 935 nm, 1060 nm, 1300 nm, 1550 nm were evaluated, as well as a composite image formed by combining 1060 nm, 1300 nm, and 1550 nm. Deep features extracted from the *fc7* layer of VGGFace network were used, resulting in a 4096-dimensional feature vector. Euclidean distance calculated between feature representations of VIS and SWIR images was used as a similarity comparison.

2.4.2 Emissive IR-to-visible *FR*

MWIR-VIS

Appearance difference between MWIR and VIS is more pronounced than NIR and SWIR (Figure 2.2). Therefore, learning a common feature subspace for MWIR-to-VIS CFR scenario is a challenging problem. In most cases, methods developed over the years have relied on GANs to synthesize visible face images from thermal face images. Illustrative examples are shown in Figure 2.10. We report the related MWIR-to-VIS face comparison performance in Table 2.5.

DPM. One of the first successful MWIR-to-VIS approaches with deep neural network was introduced by Sarfraz *et al.* [120], termed *Deep Perceptual Mapping* (DPM). Their approach was targeted to learn a non-linear mapping between visible and thermal domains (including MWIR and LWIR) while preserving the identity information. Specifically, a feed-forward neural network

was constructed to regress densely computed features from the visible image to the corresponding thermal image. An MSE loss function was used to measure the perceptual difference between the visible and thermal images. Further, a regularization term with Frobenius norm of the projection matrix was also introduced.

TV-GAN. Zhang *et al.* [150] proposed a *Thermal-to-Visible Generative Adversarial Network* (TV-GAN) that can synthesize visible images from their corresponding thermal images while maintaining identity information during the reconstruction. To preserve the identity information, a multi-task discriminator was designed. This discriminator served two different purposes. Firstly, one output of this discriminator was used to differentiate whether the generated samples were "real" or "fake". Secondly, the other output of this discriminator was used to perform identity classification in the context of supervised learning, given that the label identities are provided. Note that this identity classification was regarded as a closed-set FR. That means the identity associated with a given sample had already been seen in the training dataset. The proposed TV-GAN is similar to the auxiliary classifier GAN (or AC-GAN). Experiments were conducted on IRIS dataset [132] and compared against three baselines approaches.

SG-GAN. Chen *et al.* [21] proposed the use of *Semantic-Guided Generative Adversarial Network* (SG-GAN) to automatically synthesize visible face images from their thermal counterparts. Specifically, semantic labels, extracted by a face parsing network were used to compute a semantic loss function to regularize the adversarial network during training. These semantic cues denoted high-level facial component information associated with each pixel. Further, an identity extraction network was leveraged to generate multi-scale features to compute an identity loss function. To achieve photo-realistic results, a perceptual loss function was introduced during network training to ensure that the synthesized visible face was perceptually similar to the target visible face image. Experiments involving *PCSO* [81] face dataset showed that the proposed method achieved promising results in both face synthesis and CFR (see Figure 2.10).

Summary of MWIR-to-VIS Face Comparison. MWIR-to-VIS CFR scenario has been scarsely studied in scientific literature. Partly this is the case due to lack of public benchmark datasets. Interestingly, early research on comparing MWIR-to-VIS face images adopted a neural network [120], formulated by sequential operations of fully connected layer and non-activation layer. This simple feed-forward neural network does not involve the use of convolution units and can be represented as

$$H(x) = h^N = g(W^N h^{N-1} + b^N), \tag{2.15}$$

where, W is the projection matrix, h is the hidden layer and b is the bias. Basically, the output of current hidden layer is a multiplication of project matrix with the output of previous hidden layer. g is a non-activation function to ensure that the mapping is non-linear. The objective function of DPM was formulated

as

$$\arg\min_{W,b} \mathcal{L} = \frac{1}{M}\sum_{i=1}^{M}(\bar{x}_i - t_i)^2 + \frac{\lambda}{N}\sum_{i=1}^{N}\left(\|W^k\|_F^2 + \|b^k\|_2^2\right). \tag{2.16}$$

Here, the first term is the mean square error between the feature vectors of visible $\bar{x}_i$ and thermal t_i image. The second term serves as a regularizer on the weight matrix W and bias b. Although the DPM showed promising results on comparing SWIR against VIS images, the performance was still far from satisfactory.

Recent developments of neural networks, specifically in GANs, have pushed the frontiers of SWIR-to-VIS CFR scenario. Both TV-GAN [150] and SG-GAN [21] are inspired by Pix2Pix [68], which involves the use of adversarial loss and $\mathcal{L}_1$ loss to regularize the image synthesis process. The $\mathcal{L}_1$ loss was used to measure the per-pixel difference between synthesized visible face image and target visible face image. Compared to Pix2Pix, TV-GAN modified the output of the discriminator to perform a closed-set identification. Therefore, a new identity loss was introduced in TV-GAN. SG-GAN, on the other hand, adds more loss functions to the Pix2Pix, including perceptual loss, identity loss and semantic loss. An ablation study was used to demonstrate the individual functionalities of the loss functions (see Figure 2.10).

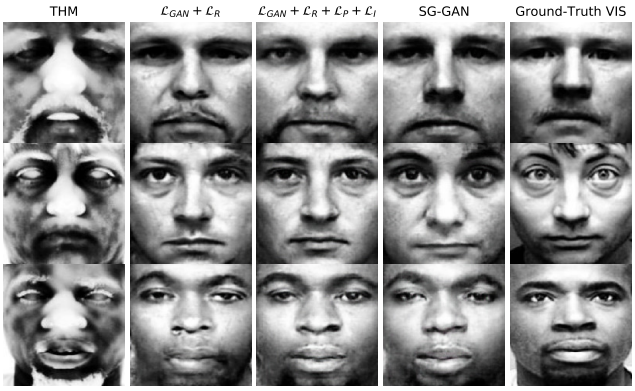


FIGURE 2.10: **Synthesizing VIS face images** from THM images on the PCSO dataset. Compared to the use of the "$\mathcal{L}_{GAN} + \mathcal{L}_R + \mathcal{L}_P + \mathcal{L}_I$" loss function, the output of SG-GAN is semantically more close to the ground-truth VIS image especially around the salient facial regions.

LWIR-VIS

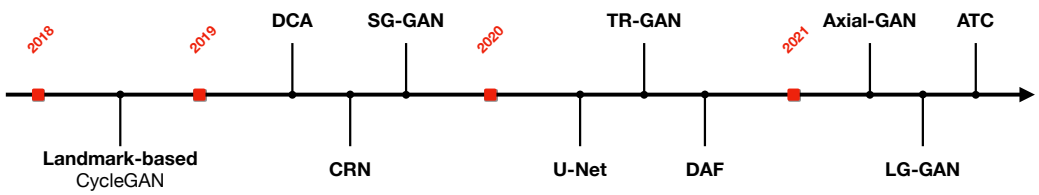


FIGURE 2.11: **Timeline of developments in algorithms:** LWIR-to-VIS CFR.

TABLE 2.6: Rank-1 Accuracy and Verification Rate on different dataset for the **LWIR-VIS** face Comparison.

Authors	**Methods**	**Dataset**	**Performance** %
Sarfraz *et al.* [120]	DPM	UND X1 [107]	83.73 @Rank-1
		Carl [37]	71.00 @Rank-1
		NVESD [16]	97.33 @Rank-1
		Polarimetric thermal [52]-*ext. - Therm. Probe*	84.00 @Rank-1
Zhang *et al.* [147]	GAN-VFS	Polarimetric thermal [52] - *Therm. Probe*	85.61 @Rank-1 - 79.30 @AUC - 27.34 @EER
		ARL-VTF [114]	97.94 @AUC - 08.14 @EER
Wang *et al.* [137]	CycleGAN [155]	private database	88.90 @Rank-1 - 98.80 @Rank-3 - 99.80 @Rank-5
	CycleGAN-Based	private database	91.60 @Rank-1 - 99.30 @Rank-3 - 99.90 @Rank-5
Iranmanesh *et al.* [67]	CpDCNN	Polarimetric thermal [52] - *Therm. Probe*	88.57 @Rank-1
Di *et al.* [33]	AP-GAN	Polarimetric thermal [52] - *Therm. Probe*	84.16 @AUC - 23.90 @EER
Kantarcı *et al.* [76]	DCA	UND X1 [107]	87.20 @Rank-1
		EURECOM [76]	88.33 @Rank-1
		Carl [37]	85.00 @Rank-1
Mallat *et al.* [100]	CRN	EURECOM [76]	82.00 @LightCNN
Iranmanesh *et al.* [65]	AGC-GAN	Polarimetric thermal [52] - *Therm. Probe*	89.25 @Rank-1
Di *et al.* [31]	SAGAN	Polarimetric thermal [52] - *Therm. Probe*	91.49 @AUC - 15.45 @EER
		ARL-VTF [114]	99.28 @AUC - 03.97 @EER
Chen *et al.* [21]	SG-GAN	Polarimetric thermal [52] - *Therm. Probe*	93.08 @AUC - 14.24 @EER
Chatterjee *et al.* [19]	U-Net	Nagoya University	69.60 @Rank-1
Kezebou *et al.* [78]	TR-GAN	TUFTS [110]	80.70 @Resnet50 - 88.65 @VGG16
Iranmanesh *et al.* [66]	CpGAN	UND X1 [107]	76.40 @Rank-1
		NVESD [16]	93.90 @Rank-1
		Polarimetric thermal [52] - *Therm. Probe*	89.05 @Rank-1
Di *et al.* [34]	Multi-AP-GAN	Polarimetric thermal [52] - *Therm. Probe*	90.99 @AUC - 17.81 @EER
Fondje *et al.* [40]	DAF	Polarimetric thermal [52]-*ext. - Therm. Probe*	94.20 @Rank-1
		ARL-VTF [114]	99.76 @AUC - 2.30 @EER
Immidisetti *et al.* [63]	Axial-GAN	ARL-VTF [114]	94.4 @AUC - 12.38 @EER
Anghelone *et al.* [7]	LG-GAN	Polarimetric thermal [52] - *Therm. Probe*	93.99 @AUC - 13.02 @EER
		ARL-VTF [114]	96.96 @AUC - 5.94 @EER
Peri *et al.* [112]	Pix2Pix-ATC	ARL-VTF [114]	91.3 @AUC - 16 @EER
	CycleGAN-ATC	ARL-VTF [114]	96.8 @AUC - 9.2 @EER
	CUT-ATC	ARL-VTF [114]	97.7 @AUC - 6.9 @EER
		TUFT [110]	87.4 @AUC - 21.3 @EER
Chen *et al.* [20]	AG-GAN	ARL-VTF [114]	99.26 @AUC - 4.30 @EER
		SF [2]	90.53 @AUC - 17.20 @EER
Anghelone *et al.* [6]	ANYRES	ARL-VTF [114]	99.88 @AUC - 1.26 @EER
		SF [2]	91.39 @AUC - 15.87 @EER
		EURECOM [76]	93.65 @AUC - 14.05 @EER
		TUFTS [110]	83.09 @AUC - 24.53 @EER

Face images captured under MWIR and LWIR tend to look very similar (Figure 2.2). Therefore, algorithms developed to address LWIR-to-VIS closely resembles that of MWIR-to-VIS. Most developed algorithms for LWIR-to-VIS CFR scenario were based on GANs. Here, we give a brief summary of recent representative works for LWIR-to-VIS comparison, related performance are reported in Table 2.6 and we illustrate the timeline of developments in Figure 2.11.

Landmark-based CycleGAN. Wang *et al.* [137] proposed a detection network to extract facial landmarks from visible faces and use that to guide the generative network to preserve geometric shapes. Their network was based on *CycleGAN* [155], targeted to perform translation between the thermal and visible face images. The detection network extracted 68-landmarks from visible faces, constructing the shape loss function. To perform the experiment, they established a new database (available online), including 33 subjects with 792 thermal-visible pair images.

DCA. Kantarcı *et al.* [76] presented a *Deep Convolutionnal Autoencoder* (DCA), specifically based on the *U-Net* [119] architecture, including Up Convolution and Difference of Gaussians (DoG), which was modified it for the purpose to learn the non-linear mapping between visible and thermal face images for CFR. In the meantime, they showed that applying preprocessing and alignment would help to close the gap between these two domains and was able to improve the performance further. The proposed approach was extensively tested on three publicly available thermal-visible datasets: Carl [37], UND-X1 [107] and EURECOM [98] (also known as VIS-TH dataset). In addition, authors have manually annotated 6 facial landmarks on the Carl and EURECOM datasets in order to align the faces and assess alignment's impact on the performance (we observe that alignment increased the performance by around 2%). They conduct experiments with three different setups: the first two experiments investigated the effect of the decoder method, DoG filter were not applied to images as preprocessing step. For the third experiment, the effect of the DoG filter was tested on the Up Convolutionnal Autoencoder Model, proposed in this work.

CRN. Mallat *et al.* [100] proposed a solution based on *Cascaded Refinement Network* (CRN). Their approach did not require a large amount of training data, owing to a limited number of training parameters. CRN is a type of CNN that consists of inter-connected refinement modules, which gradually process the image synthesis from lowest resolution (4×4) to the highest resolution (128×128). The refine module includes three different layers, namely input, intermediate and output layers. A contextual loss function was used to train the CRN. The motivation to choose the contextual loss was based on two conditions: (a) robustness to roughly aligned THM-VIS pairs; and (b) invariant to outlier at the pixel level in the context of per-pixel loss. FR experiments on the EURECOM [76] dataset showed that the proposed method achieves better performance than TV-GAN. In addition, they performed CFR using their synthesized faces (facial variation set to Neutral) with two systems, namely *OpenFace* and *LightCNN*.

U-Net. Chatterjee and Chu [19] presented a U-Net architecture with a

residual network backbone to generate visible face images from thermal face images. The U-Net was modified by using residual blocks with skip-connections instead of the normal convolutional layers as the basic building components. Further, a pixel shuffle upsampling was introduced to replace the transposed convolution layers in the decoder part. A weighted combinations of two different loss functions was used to train the proposed network, namely mean squared loss and perceptual loss. The proposed method was evaluated on thermal face dataset from *Nagoya University*, which contained 900 thermal images and 900 visible images captured simultaneously. A Rank-1 accuracy of 69.60% was achieved.

TR-GAN. Kezebou *et al.* [78] proposed a *Thermal to RGB Generative Adversarial Network* (TR-GAN) to automatically synthesize visible face images captured in the thermal domain. The TR-GAN employed an architecture similar to U-Net with cascade residual blocks for the generator. Basically, it replaced the Resnet blocks in the CycleGAN with cascaded-in-cascaded residual blocks. This ensured that the generator synthesizes images with consistent global and local structural information. For the discriminator, the TR-GAN used the same discriminator network architecture as CyleGAN. A pretrained VGG-Face recognition model was used to perform the face comparison after the thermal to visible image translation. The experiments were conducted on TUFTS face dataset and compared against TV-GAN, Pix2PixHD and CycleGAN.

DAF. Fondje *et al.* [40] proposed a *Domain Adaptation Framework* (DAF) with a new feature mapping sub-network including (a) a *cross-domain identification*-loss function for effective “co-registration" and “synchronization" and (b) a *domain invariance*-loss function for cross-domain regularization. Within each domain, the feature representations were acquired using a truncated pretrained CNN (based on VGG16 or ResNet50). The authors were interested in finding the optimal depth-level of the network while preserving the meaningful discriminative information shared by both the visible and thermal spectra. DAF is then embedded in their proposed *Residual Spectral Transform* (RST), which is a residual block. The key role of RST is to transform the features with the help of three particular 1x1 convolutions while enhancing the discriminability. Experiments were conducted using three datasets (considering only thermal probes): the polarimetric thermal dataset [52], the polarimetrics thermal dataset extended to 111 subjects [148],[52]-*ext.* and a private dataset containing paired visible/thermal faces of 126 subjetcs. DAF was mainly compared to DPM [120] and shows significant improvement.

Axial-GAN. Immidisetti *et al.* [63] proposed Axial-Generative Adversarial Network (Axial-GAN) to synthesize high-resolution visible images from low-resolution thermal counterparts. The proposed GAN framework designed an axial-attention layer with transformer to model long-range dependencies to facilitate long-distance face matching. Their work can simultaneously address face hallucination and translation for thermal-to-visible face matching. Evaluations on ARL-VTF thermal and multi-modal polarimetric thermal face recognition datasets obtained promising performances compared to SAGAN and HiFaceGAN.

ATC. Peri *et al.* [112] presented a comprehensive study on synthesis-based approach for thermal-to-visible face verification. They explored Pix2Pix, CycleGAN and Contrastive Unpaired Translation (CUT) off-the-shelf domain adaptation algorithms for the task of generating realistic synthetic samples. They further demonstrated that additional custom loss such as Pixel wise correspondence loss and identity loss are instrumental for addressing the thermal-to-visible face verification. Authors highlighted the impact of the face alignment and encompass their method named ATC with above loss functions in a end-to-end thermal-to-visible system comprising ATC: Alignment, spectrum Translation and Classification steps.

Summary of LWIR-to-VIS Face Comparison. LWIR-to-VIS CFR scenario resembles MWIR-to-VIS CFR scenario as face images captured in LWIR and MWIR spectral bands are almost indistinguishable. In the previous works of MWIR-to-VIS comparison, algorithms developed based on Pix2Pix have been used to address the challenge [150, 21]. The Pix2Pix model was designed only to learn forward mapping from one domain to another, while CycleGAN was streamlined to learn both the forward and inverse mappings simultaneously using cycle-consistency loss. In the work of [137] (method entitled Landmark-based CycleGAN), CycleGAN was used to learn such bi-directional mappings for LWIR and VIS face images. The objective function of CycleGAN is defined as

$$\begin{aligned}\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{GAN}(G, D_Y, X, Y) \\ + \mathcal{L}_{GAN}(F, D_X, Y, X) \\ + \lambda \mathcal{L}_{cyc}(G, F).\end{aligned} \tag{2.17}$$

Here, G is used to learn the forward mapping from X to Y, *i.e.*, $G : \{X \rightarrow Y\}$. The corresponding discriminator is D_Y. Therefore, $\mathcal{L}_{GAN}(G, D_Y, X, Y)$ can be described by

$$\begin{aligned}\mathcal{L}_{GAN}(G, D_Y, X, Y) = \mathbb{E}_{y \sim p_{data}(y)}[log(D_Y(y))] \\ + \mathbb{E}_{x \sim p_{data}(x)}[log(1 - D_Y(G(x)))].\end{aligned} \tag{2.18}$$

Similarly, F is used to learn the inverse mapping from Y to X, *i.e.*, $F : \{Y \rightarrow X\}$. The corresponding discriminator is D_X. $\mathcal{L}_{GAN}(F, D_X, Y, X)$ can be described by

$$\begin{aligned}\mathcal{L}_{GAN}(F, D_X, Y, X) = \mathbb{E}_{x \sim p_{data}(x)}[log(D_X(x))] \\ + \mathbb{E}_{y \sim p_{data}(y)}[log(1 - D_X(G(y)))].\end{aligned} \tag{2.19}$$

In addition to adversarial losses of $\mathcal{L}_{GAN}(G, D_Y, X, Y)$ and $\mathcal{L}_{GAN}(F, D_X, Y, X)$, a cycle consistency loss was also introduced to ensure that an input image x in domain X can be successfully reconstructed after going through both the forward mapping G and the inverse mapping F, *i.e.*, $x \rightarrow G(x) \rightarrow F(G(x)) \rightarrow$

$\bar{x}$. The cycle consistency loss is described by

$$\begin{aligned}\mathcal{L}_{cyc}(G,F) = \mathbb{E}_{x\sim p_{data}(x)}\|F(G(x)) - x\|_1 \\ + \mathbb{E}_{y\sim p_{data}(y)}\|G(F(y)) - y\|_1.\end{aligned} \tag{2.20}$$

Using CycleGAN alone allowed the generation of visually pleasant visible images from thermal images. However, adding more constraints to restrict the mappings is envisioned further improve the synthesis quality. In [137], a 68-point landmark detection network was used to extract the landmarks from the visible images. This can assist the CycleGAN to preserve the geometrical shapes of the reconstructed samples. The shape loss was defined as

$$\mathcal{L}_{shape} = \frac{1}{68}[(y_s - G(x)_s)^2 + (y_s - F(G(y))_s)^2], \tag{2.21}$$

where, x is a thermal image and $G(x)$ is a synthesize visible image from generator G. y is a visible image and $F(G(y))$ is a reconstructed visible image.

2.5 Summary

In this Chapter we provided a comprehensive overview of CFR methods. The formalization of CFR is instrumental in comparing existing methods, as well as in addressing the challenge of cross-domain feature comparison. We discussed spectral bands and introduced general components of a CFR system, including face processing, deep feature extraction (CNNs, GANs and loss functions), face comparison and biometric evaluation metrics. This study has brought to the fore crucial aspects of a system, thus supporting the design of our pipeline as an end-to-end method. We revisited then prior works and determined strengths and limitations. In the context of the invisible spectrum, we specifically investigated the infrared spectrum ranging from near infrared (NIR) to the more challenging bands such as longwave-thermal infrared (LWIR). In particular, we reviewed comprehensively proposed algorithms, as well as results and associated insights. We emphasized on advances of *deep learning* in addressing CFR and techniques targets to bridge the modality gap.

Chapter 3

Multi-spectral Face Dataset

The collection of a comprehensive and diverse face dataset is crucial for research in the field of CFR. The most significant research achievements made over the past decade have heavily relied on the availability of new and diverse datasets, as well as the design of benchmarks for evaluating CFR algorithms. Hence, the advancement of CFR is heavily dependent on the availability of training data, and as deep learning approaches are known to be *data-hungry*, the collection and availability of training samples are pivotal elements in the advancement of this field.

This Chapter introduces the databases that are dedicated to train CFR. Section 3.1 reviews existing datasets that are publicly available for CFR research, grouping them according to the spectral bands in which the images were acquired. Our aim is to provide a comprehensive overview of the current state of the art in terms of datasets and to highlight their strengths and limitations. Section 3.2 provides insight into the collection of our in-house multi-spectral resources. Particular efforts are made to address deficiencies in variability in terms of *unconstrained* environments data, in which other datasets currently lack. Finally, Section 3.3 brings a preliminary evaluation on comparing thermal probe face images against a visible gallery of faces. This study aims to deepen our understanding regarding robustness and accuracy over variations such as pose, expression, occlusion, poor image quality, and distance of capture.

3.1 Existing multi-spectral face databases

CFR is a challenging task that requires large-scale, heterogeneous facial datasets. However, the availability of public face datasets across spectra is still limited, particularly in terms of variability. In the following Section, a comprehensive review of the main characteristics of datasets designed for CFR will be presented and summarized in Table 3.1. These datasets will be grouped according to the infrared spectral bands in which the images were acquired. It will provide a clear understanding of the current state-of-the-art in terms of available resources for CFR research.

TABLE 3.1: Summary of the main databases of multi-spectral facial images. The variability nomenclature is denoted as follows: (P)ose, (E)xpression, (I)llumination, (G)lasses, (O)cclusion, (D)istance and (T)ime-lapse.

Spectra	Name	Subjects	Images	Variability
NIR-VIS	CASIA HFB [93]	100	5097	Neutral
	CASIA NIR-VIS 2.0 [94]	725	17580	P,E,G,D
	Oulu-CASIA NIR-VIS [152]	80	7680	E,I
	BUAA-VisNir [59]	150	2700	P,E
	LAMP-HQ [146]	573	73616	P,I
	LDHF-DB [97]	100	-	D
SWIR-VIS	PRE-TINDERS [131]	48	576	E
	TINDERS [131]	48	1255	E,D
MWIR-VIS	WSRI [26]	64	3230	E
	MILAB-VTF(B) [112]	400	-	P,D
LWIR-VIS	UND X1 [107]	241	4584	E,I,T
	ARL-MMFD [52]	60	2280	E,D
	TUFTS [109]	113	900	P,E
	VIS-TH [98]	50	2100	P,E,I,O
	SF [2]	142	7668	P,E
	ARL-VTF [114]	395	549712	P,E,G

3.1.1 NIR-VIS

One of the earliest benchmark datasets designed for CFR is the **CASIA HFB** dataset [93]. It includes 100 individuals with 4 acquisitions per subject in both NIR and VIS spectra, as well as a 3D depth image. An extension of this dataset is the **CASIA NIR-VIS 2.0** dataset [94]. It represents one of the first large NIR-VIS face datasets, containing 17,580 NIR-VIS face images of 725 subjects with variations in pose, expressions, eyeglasses, and distances. The CASIA NIR-VIS 2.0 dataset represents a significant advancement in comparison to the CASIA HFB dataset, including an increased number of subjects by a factor of three, an expanded age distribution among the subjects, and an evaluation protocol for benchmarking. As a result, the CASIA NIR-VIS 2.0 dataset has become the most commonly used benchmark dataset for the evaluation of NIR-VIS CFR scenario.

Further databases provided distinct characteristics. For instance, the **Oulu-CASIA NIR-VIS** dataset [152] includes 80 subjects with 6 variations of facial expressions such as: anger, disgust, fear, happiness, sadness, and surprise. Moreover, resulting from several acquisitions under various environments, the facial images exhibit slight differences in illumination conditions. Unlike prior datasets mentioned, Oulu-CASIA NIR-VIS provides 48 VIS and 48 NIR samples per subject, thus being much desirable for learning changes in facial expression across spectrum.

Exploration of *domain adaptation* techniques is particularly relevant for CFR (see Section 2.4), because of the low discrepancy in modality between the near infrared and visible spectrum. In this context, the **BUAA-VisNir** dataset [59] is a valuable resource, as it contains a large number of images from a diverse set of 150 subjects, i.e. 9 VIS and 9 NIR images along with variation in poses and expressions. However, the training and testing sets exhibit larger differences making them challenging for evaluation.

To further expand the range of variability of training samples, the **Large-Scale Multi-Pose High-Quality** [146] database (LAMP-HQ) proposes a large scale database over 73,000 images, from 573 subjects, having large diversities in pose, illumination, attribute, scene, and accessory.

Currently, publicly available datasets in CFR research are relatively limited. While there are several datasets that have been widely used for benchmarking CFR algorithms, these datasets have not yet addressed the problem of CFR at long distances. The **LDHF-DB** dataset [97] offers a unique opportunity to address the challenge of CFR in unconstrained environment. It includes subjects captured at distances of 60 meters, 100 meters, and 150 meters, in both VIS and NIR spectral bands. Thus, it is highly instrumental in investigating cross-distance and cross-spectral NIR-VIS FR.

3.1.2 SWIR-VIS

In contrast to NIR imaging, collecting images with a SWIR camera is a more complex task. Although SWIR is also spectrally reflection-dominated, the spectral response requires being filtered to capture a reflected light, and thus images at the desired wavelength. This complexity can present challenges in terms of image acquisition and processing, making the collection of large-scale datasets in the SWIR spectrum more tedious to achieve.

The **PRE-TINDERS** [131] dataset is a pioneering public dataset and valuable resource for CFR research. However, it provides a relatively limited set of 484 images captured from 48 subjects, in both spectra. One of the strengths is that it contains images of subjects with neutral expressions as well as mouth expression, thus allowing the study of facial dynamics additionally to static facial features. Moreover, the close distance of acquisition and the use of a 1550 nm light source for the SWIR images ensures a high level of image quality

(see Figure 2.1 to localize the associated wavelengths in the electromagnetic spectrum). Indeed, SWIR imaging is highly sensitive to this wavelength. Nevertheless, the limited number of subjects and the single session acquisition may limit the generalizability of any findings.

To bring variability in distance acquisition as well as to allow for a more comprehensive study of cross-distance recognition in the SWIR spectrum, the **TINDERS** [131] dataset is introduced as extension of the PRE-TINDERS dataset. TINDERS also depicts 48 subjects over neutral and talking expressions but includes additional SWIR images captured at different distances, *i.e.,* 50 and 106 meters. A total of 478 images are available in the SWIR band at different distances. In contrast, visible images were collected at a short distance in two other sessions, entailing neutral expression over 288 images.

Availability of public datasets for SWIR-VIS FR are limited. The PRE-TINDERS and TINDERS datasets, although providing valuable resources for researchers in this field, are currently among the few available options. This highlights the need for further efforts to be made in the collection and dissemination of such datasets to support the advancement of CFR methods. Despite these challenges, the use of SWIR imaging in conjunction with visible imaging, has the potential to provide additional information and improve the performance of FR systems.

3.1.3 MWIR-VIS

The high cost of MWIR sensors and the range of action scoped to military applications entail the availability of public MWIR-VIS cross-spectral datasets. Hence, only few datasets have been made available for research purposes.

Among those, the **WSRI** dataset [26] offers a collection of VIS and MWIR images from 64 subjects, with 25 images per individual including different facial expressions. This dataset presents several strengths such as a high number of images and a wide range of facial expressions. However, it also has limitations such as the small number of subjects and the fact that the images were captured in controlled conditions.

Further extending variability to unconstrained settings, the **MILAB-VTF(B)** [112] dataset is presented at the time of writing this dissertation as the largest collection of long-range unconstrained and (unsynchronized)-paired MWIR-VIS face images overs 400 subjects. It represents a significant strength in terms of diversity of subjects and pose capture settings, including both indoor and outdoor acquisitions at distances ranging from 100 to 400 meters. However, it should be noted that the images were not captured simultaneously, which may present a limitation in terms of accurately evaluating the performance of CFR methods when *domain translation* methods are developed (see Section 2.4).

3.1.4 LWIR-VIS

LWIR imaging has gained increasing attention in recent years due to the more affordable costs of sensors, allowing thermal acquisition with even better quality sensors.

The **UND X1** [107] dataset offers a diverse range of data, including 241 subjects with different variations in lighting, expression and especially designed to study time lapse impact on thermal images. However, the low resolution and noise present in the LWIR imagery can pose a challenge for allowing deep models to learn pertinent facial features. Furthermore, evaluation of methods are impeded by the unbalanced training/testing protocol. The training set is composed of 159 subjects, with only one image per subject. On the other hand, the test set is contained of the 82 remaining subjects with multiple images per subject. As a result, deep models have only few samples to learn and in contrast a quantity of data to test, thus lacking generalisation.

The **ARL-MMFD** [52] dataset is a multi-spectral dataset containing pairs of images captured at different wavelengths, including the visible, thermal, and polarimetric spectra. One of the main advantages of this dataset is that it offers images captured at different distances, 2.5m , 5m and 7.5m, respectively. Furthermore, the set is directly provided with aligned face images. However, subjects are presented only with neutral expression without variation in pose, expression or occlusion. Moreover, the dataset is limited in terms of number of images and identities, as only 60 subjects, later extended to 111 subjects, have been collected. As a result, there is a lack of subjects diversity. Additionally, the training and testing sets are poor in terms of data, making challenging to evaluate the performance of FR algorithms using this dataset.

The **TUFTS** dataset [109] provides a large number of images, namely over 10,000 images from 113 individuals, acquired in various modalities including visible, thermal, and also near infrared images. Benefits from this collection are the simultaneous acquisitions across devices, along with a set of scenarios ranging from pose variations to expressions with sunglasses.

Unlike the TUFTS dataset, the **VIS-TH** dataset [98] extends the range of variations by considering facial expressions, poses, different illumination conditions, several occlusions which also includes eye and mouth occluded by subject's hands. In total, there are 2100 images that belong to 50 subjects.

Pose variation samples are valuable in training deep models dedicated to unconstrained FR. The **SF** [2] dataset contains a rich set of images simultaneously captured under VIS and LWIR spectra. The collection includes 142 subjects and combines variation in speaking expressions along with 9 extreme facial poses (yaw and pitch). These features enable the study of real-world scenarios where faces are captured under different conditions.

The aforementioned datasets present a wide range of acquisitions but have a significant limitation in terms of number of subjects. This results in unbalanced training and testing sets. However, with a total of 395 subjects and over half a million images captured simultaneously under LWIR and VIS spectra, the **ARL-VTF** [114] dataset provides a vast array of variations in expression, pose, and eyewear, making it one of the most recent comprehensive datasets publicly available. Additionally, it is endowed with annotations, metadata, and standardized protocols for fair evaluation, making it a valuable resource for researchers in the field. Nevertheless, variations in pose and expression are restricted to only yaw movements and changes in the mouth. Furthermore, there are no considerations regarding long range distance acquisition, thus making this dataset less adapted for the use in unconstrained environments.

To conclude, existing datasets provide a good starting point for CFR research, however they are limited in terms of variability, especially in unconstrained scenarios and long-distance acquisitions. Therefore, there is still a need for more diverse and large datasets to advance the research in this field. Moreover, despite the increasing resource availability and decreasing costs sensors, there are still challenges to be addressed in terms of infrared data collection and image normalization, as lighting conditions and temperature variations can greatly affect the quality of the images.

3.2 BYDB: Beyond the Visible Database

Despite the availability of several public multi-spectral paired face datasets, there remains significant limitations involving a lack of diversity in terms of subjects, limited variations in pose, expression, occlusion and of the distance acquisition, as well as a lack of simultaneous capture across several spectra and a poor image quality. Additionally, many datasets are restricted in terms of the number of subjects and images, where the training and testing protocols are often unbalanced.

These limitation serve as motivation for the development of our own dataset, referred to as *Beyond the Visible Database* (BYDB), described in this Section. Specifically, BYDB is a multi-spectral collection of facial images and videos, endowed with annotations and metadata. In particular, faces are captured simultaneously across four electromagnetic spectra, including the VIS, NIR, SWIR and LWIR spectral bands. BYDB differs from other datasets in many aspects. With a wide range of variations in terms of *acquisition distances, illuminations, poses, expressions, occlusions, makeup and time-lapse*, it offers a vast array of data for training and assessing CFR systems in real-world scenarios, including unconstrained settings. This dataset is a valuable resource for Thales, including researchers and practitioners from the industry working in the field of FR, particularly in the areas of cross-spectral and long-distance FR. With $1,476,292$ of images and a large number of subjects, BYDB is an unprecedented dataset that will enable new breakthroughs in the field. We note

that BYDB was collected within the framework of a CIFRE-fellowship thesis, and is hence classified private and internal to Thales.

3.2.1 Face sensor suite

Cameras

TABLE 3.2: Camera specification pertaining to the BYDB data collection.

Camera	Spectrum	Resolution (px)	Sensor
Basler VIS	VIS	1920×1200	CMOS
Basler NIR	NIR	1280×1024	CMOS
NIT WiDy 640	SWIR	640×512	InGaAs
FLIR A700	LWIR	640×480	Uncooled Microbolometer

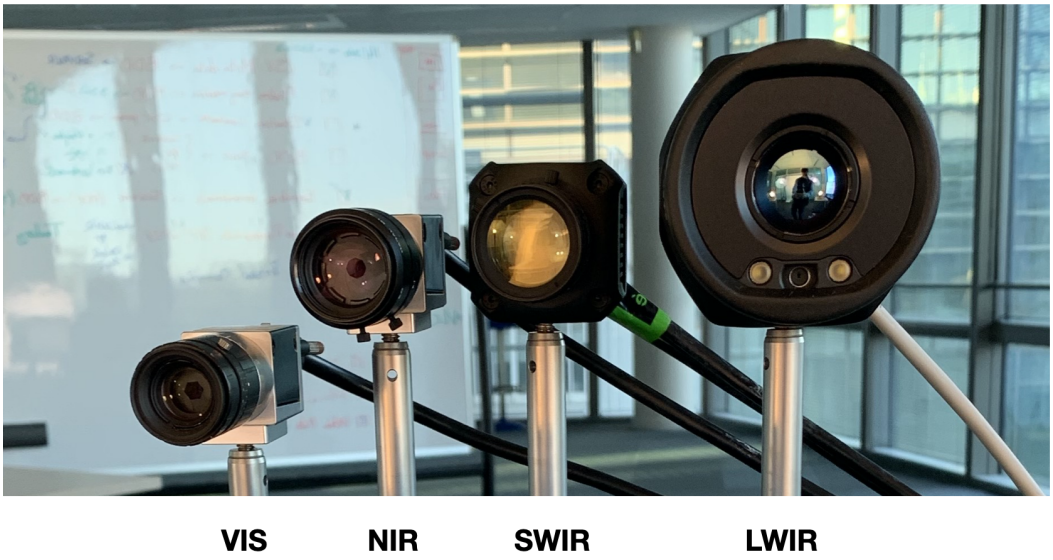


FIGURE 3.1: Sensor suite used for the collection of BYDB dataset. The acquisition setup comprises an array of four cameras responsive to the visible (VIS), near-infrared (NIR), short-wave infrared (SWIR) ans long-wave infrared (LWIR) spectra, respectively.

To acquire multi-spectral facial images, four different sensors were selected to cover specific sensitive wavelengths of the electromagnetic spectrum, in both visible and infrared spectra. The hardware specification of these devices are listed in Table 3.2, while the complete sensor suite is shown in Figure 3.1. This combination of cameras allowed us to acquire a wide range of multi-spectral data which comprises the VIS, NIR, SWIR and LWIR modalities. The MWIR spectrum is however not included in the collection for cost and practical reasons. At the time of writing this dissertation, the above infrared SWIR and LWIR sensors are considered as the state-of-the art technology, offering a high resolute image.

Calibration and synchronization

Technologies related to the cameras are different, as can be seen in the *sensor* column of Table 3.2. The VIS, NIR and SWIR cameras are embedded with a Global Shutter, while the LWIR camera operates with a Rolling Shutter. Thus capturing details of faces across multiple modalities in a time-synchronized manner requires a specific calibration for each sensor. The calibration needs to be conducted before each new volunteer acquisition and involves a first step of adjusting the focus for all cameras, then ensuring that the exposure is appropriate, the white balance is neutral, and there are no anomalies present on the display screen.

One of the major concerns in designing the acquisition setup is ensuring that the four cameras are able to capture the same information from the same scene without any difference in terms of angle of view. Since the optical field and the position of the sensors are different in space, there is a need to constrain the gap of parallax. Therefore, a custom plate has been designed to fix the four cameras in order to ensure that the fields of view of the lenses are stable and fully capturing the face of the person. This operation required a careful alignment and calibration of the cameras to eliminate any discrepancies in the captured images. Note that the later fusion of captured multi-spectral images will allow transferring annotations from one spectrum to another.

Finally, in order to achieve simultaneous acquisitions, it is necessary to set the same rate of frame rate for all sensors. A custom software adapted to our needs was then implemented in order to interface with the specific software development kits (SDKs) provided by each camera's manufacturer. This software allows us to manage the four streams of information and to perform a synchronized launch of images or video acquisitions. Additionally, it also facilitate the management and analysis of the collected data by consolidating it into a single repository including separated spectrum channels sub-folders for more efficient analysis. It allows a more streamlined process, minimizing the need for manual data sorting and organization. It also enables the application of various images processing techniques to enhance the quality and utility of the data. Furthermore, by incorporating automated segmentation and data sorting algorithms, we were able to further enhance the organization and analysis of our large scale collected data.

Acquisition area

The layout of the acquisition area is shown in Figure 3.2. The studio is located in a dark room where the temperature is set to 20°C. This is done to eliminate any external lighting or temperature fluctuations that could affect the accuracy of the data collected. Therefore, it provides a consistent and standardized environment during all the campaign collection.

A volunteer sits on a chair placed one meter away from a thermally neutral white background. In this configuration, the volunteer is additionally surrounded by projectors to evenly illuminate the face and minimize shadows. Furthermore, the inclusion of infrared projectors enables the acquisition of NIR images in complete darkness.

FIGURE 3.2: A glimpse into the acquisition area setup during the BYDB database collection, showcasing the sensor suite (cameras), strategically placed projectors and the thermally neutral white background.

To capture subjects at different distances, the acquisition platform with the cameras is moved instead of moving the volunteer. This setup allows precise control over the lighting and positioning of the subject, ensuring that high-quality images are captured every time.

3.2.2 Acquisition protocol

Internal campaign and legal form

The data collection was successfully completed with legal clearance. It has been carried out with three internal Thales campaigns held in three locations in France. To ensure consistency during the campaign, the same standardized protocol of acquisition and studio area conditions were maintained throughout the entire campaign.

Before participating in the data collection project, volunteers were required to sign a consent form. Consent forms have been reviewed and approved by the legal department of Thales DIS. It is important to note that the privacy and rights of the participants were of the utmost importance and every measure was taken to ensure that their images have been used ethically and responsibly. Additionally, all information were de-identified and processed in a secure environment to protect the privacy of the volunteers. This was a crucial step in ensuring that our data collection efforts were compliant with legal and ethical guidelines of the GDPR.

Methodology

The collection protocol is comprised of three distinct types of *enrollment sessions*, which are conducted only once by the volunteer. One session is mandatory for all volunteers and serves as *basic (1) acquisition* for the collection process. Other two additional sessions are optional and designed as supplementary contents to the basic first session, offering additional data collection through *multi-distance (2)* and *technical (3)* acquisitions. Time-lapse is also considered in the protocol, in which volunteers were enthusiast to re-run the basic session few weeks later.

Finally, all three types of enrollment result in eight possible session choices which are summarized in Table 3.4. In particular, Table 3.3 reports the range of distance denoted as R1, R2 and R3, respectively fixed by the protocol.

The acquisition process involves short synchronized video recordings of 5 seconds in which the participant performs the requested actions. Each video is then sampled into a given number of images, depending on the action and whether it is *static* or *dynamic*. The number of images depends on the action. If the action is dynamic, more images were selected in order to capture a wide range of motions.

TABLE 3.3: Notation denoting the selected range distance expressed in meters (m) in the BYDB database collection.

Range notation	**Distance (m)**
R1	2
R2	5
R3	7

TABLE 3.4: Description of enrollment session options. The protocol incorporates three type of session.

Session	**Enrollment**	**Description**
Basic (1)	Choice 1	Acquisition at range R1, subject without wearing eyeglasses
	Choice 2	Acquisition at range R1, subject with wearing eyeglasses
Multi-distance (2)	Choice 3	(1) + Acquisition at range R2 or R3
	Choice 4	(1) + Acquisition at range R2 and R3
Technical (3)	Choice 5	(1) + Makeup
	Choice 6	(1) + Sport
	Choice 7	Choice 3 + Makeup
	Choice 8	Choice 3 + Sport

Scenario

BYDB comprises a wide range of variations in which some of them are depicted in Figure 3.6, these variations are referred to as scenarios and can be *simple* or

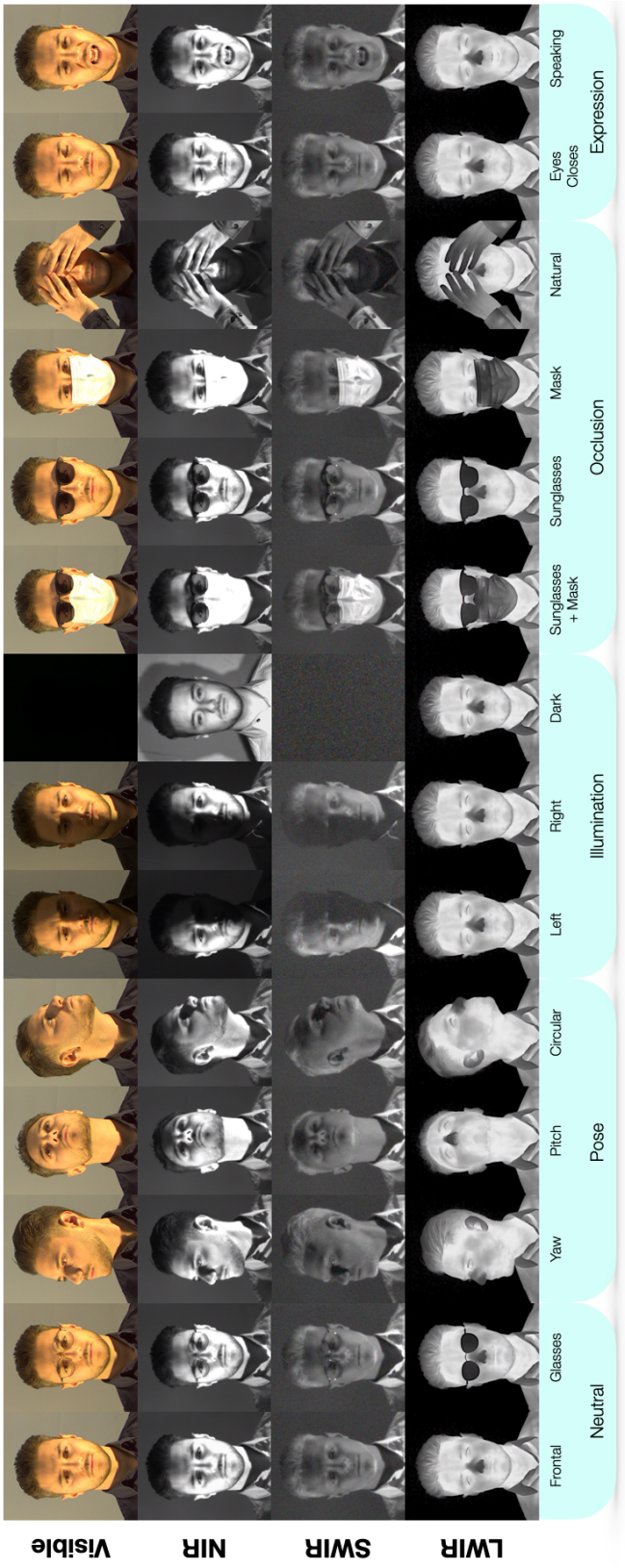


FIGURE 3.3: Illustration of the BYDB database depicting various scenarios captured simultaneously under the visible (VIS), near infrared (NIR), short-wave infrared (SWIR) and long-wave infrared (LWIR) spectral bands.

combined with each other, with actions that can be either static or dynamic. Hence, faces are recorded under following scenarios.

- **Neutral.** Baseline frontal face acquisition, including neutral facial expression with the gaze fixed on the target.
- **Illumination.** Four distinct illumination conditions are considered, involving total darkness, single illumination from the left or right and dual illumination from both left and right sources.
- **Pose.** Pose is expressed by head movement, which can by horizontal, vertical or circular, resulting in different yaw and pitch.
- **Expression.** Two types of facial expressions are considered, namely lip movements where the subject exaggerates the pronunciation of the alphabet, and closing the eyes during the acquisition sequence.
- **Occlusion.** Facial areas are hidden either by accessories such as a surgical mask, glasses or both, or by hand.
- **Distance.** Distance of acquisition differs, from 2 to 7 meters.
- **Makeup.** Military or sporting event, makeup is drawn in the face.
- **Sport.** Physical activity, in order to raise body temperature.

The acquisition is divided into two modes: *single* scenario (S) and *combined* scenario (C). A single scenario is defined as the capture of the subject with a neutral face under illumination, pose, expression or occlusion scenario, etc. We note that, a combined scenario combines two scenarios, such as horizontal (yaw) movement with left illumination or wearing a mask in the dark. In total, 55 scenarios are considered, mixing S and C mode, also considering whether the subject is wearing eyeglasses. If this later case, few scenarios are added in order to record the face, with and without wearing the eyeglasses. This strategy ensures the collection of a wide variation of acquisitions of the same person and enables to learn any subtle facial change.

3.2.3 Annotations

Annotations are essential in developing rich resources for supervised artificial intelligence algorithms. However, existing multi-spectral face datasets usually offer limited annotations which bring the scarcity of work focusing on several computer vision applications, such as facial detection or landmark detection in the invisible spectrum.

BYDB has a distinct advantage over other databases. It is endowed with *annotations*, *metadata*, and *standardized* train-test protocols, thus allowing extended research in the field. While annotations include *facial bounding box* with corresponding *facial landmarks* across all spectra, metadata(s) incorporate key information about the physiology of the subject, in terms of demographic information, gender, age, hair color, eye color, presence of beard or mustache, and whether the subject is wearing eyeglasses.

Algorithms developed for annotating visible spectrum images fail to generalize onto invisible spectrum images, this is mainly due to the modality gap. Figure 4.2 demonstrates the experimentation from [25], where an algorithm trained with visible images is trying to predict landmarks in the LWIR images. Moreover, the lack of available annotated spectral images makes it challenging to develop specific annotating tools for each spectra. Nevertheless, benefits from the synchronous acquisition allows for fusing all facial semantic information coming from one spectrum to another. Therefore, extrapolating information across spectra is of particular pertinence in acquiring rich and varied annotations.

To extrapolate ground truth annotations from the visible spectrum to the invisible spectrum: re-scale, crop and rotation through *homography* operations are applied. Considering a pair of images showing the same semantic content, a homography H is defined as a space-transformation between a *source* image and a *target* image. The same content can be overlayed by mapping a set of (x_s, y_s) points from the source image to the corresponding set of (x_t, y_t) points in the target image. Mathematically, H is defined as a 3×3 matrix that represents the mapping between two 2D plane projections. This operation is formulated as follows.

$$\begin{bmatrix} x_s \\ y_s \\ 1 \end{bmatrix} = H \begin{bmatrix} x_t \\ y_t \\ 1 \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x_t \\ y_t \\ 1 \end{bmatrix} \tag{3.1}$$

For each subject, one random quadruplet sample containing the VIS, NIR, SWIR and LWIR sync facial images is selected. On that image, five specific points on the face are *manually* annotated in order to find correspondences in-between them. These marks depict the outer of the eyes, center of the nose, and the corner of the mouth. Based on the set of points, homographies are then computed between the VIS-NIR, VIS-SWIR, VIS-LWIR, NIR-SWIR, and NIR-LWIR pair images, serving to transfer information from the visible image or the near-infrared image into the other spectra. Note that unlike the visible spectrum, the NIR imaging provides images even in the dark.

3.2.4 BYDB, details and usage

This Section aims to provide details regarding BYDB. The database includes videos with a total of 363,073 extracted facial images captured per spectrum (i.e. VIS, NIR, SWIR and LWIR) from 369 participants, enrolled under the same acquisition protocol. BYDB has also been expanded (as an extended session) to include an additional set of 43 subjects captured in uncontrolled environment. In comparison to other datasets presented in Section 3.1, BYDB is the largest collection of multi-spectral facial images captured in time-synchronized, resulting from 412 distinct identities. BYDB is compared *w.r.t.* prior introduced multi-spectral face database in Table 3.5

The data collection was performed during three main sessions including acquisition at Meudon, La Ciotat and Gemenos respectively, in Thales office buildings in France, as well as an additional session denoted as Extension. The overall campaign subject-distribution is shown in Figure 3.4. However, the rest

BYDB	Number of Subject
Meudon	142
La Ciotat	168
Gemenos	59
Extension	43
Total	412

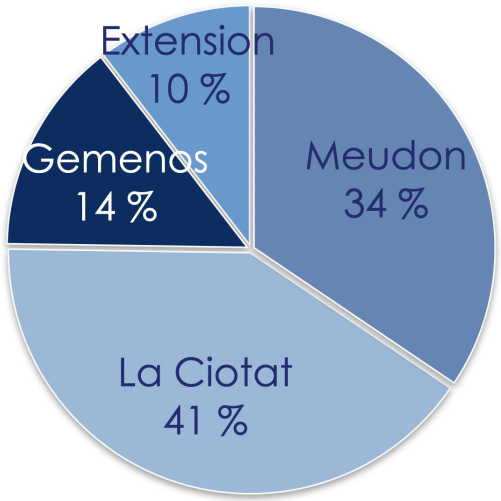


FIGURE 3.4: Subject distribution during the BYDB campaign, in the four Thales sites.

BYDB	Male	Female
Meudon	87	55
La Ciotat	96	72
Gemenos	35	24
Total	218	151

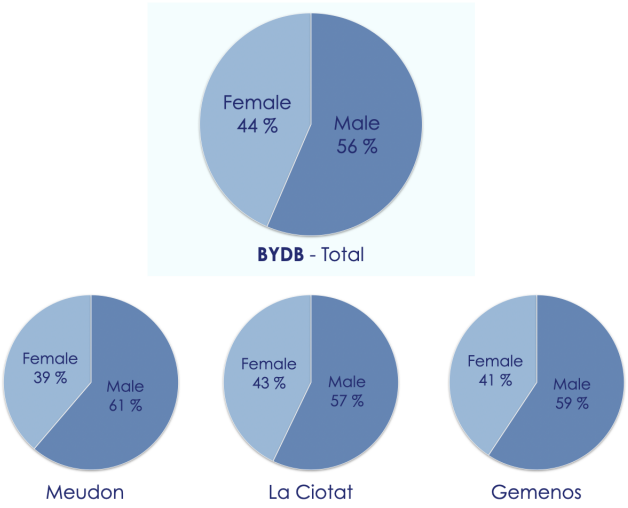


FIGURE 3.5: Gender distribution of the BYDB database.

of the dissertation will rely only on the three main regular session sets. The demographic characteristics of the dataset, which expresses *gender* and *age* distribution, are presented in Figure 3.5 and Figure 3.6, respectively.

3.3 Preliminary evaluation: comparison of thermal probe face against a visible gallery

This Section is dedicated to presenting preliminary results on comparing biometric performances when thermal probe face images are directly compared to a gallery of visible face images. The experimentation entitled *direct comparison* highlights challenges faced by the spectral modality gap and the need for dedicated algorithms to address the thermal-to-visible CFR task. Therefore,

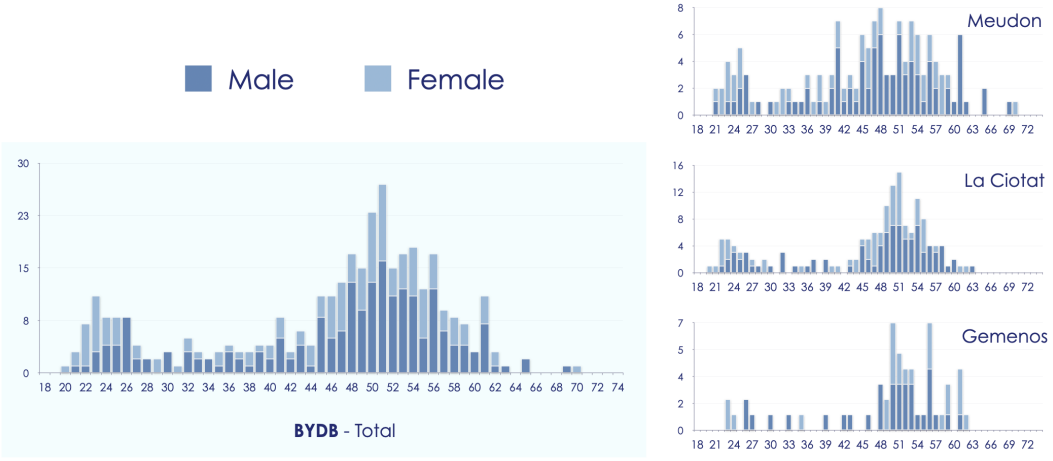


FIGURE 3.6: Age distribution of the BYDB database, with regard to the gender.

TABLE 3.5: Summary of the main databases of multi-spectral facial images in comparison to BYDB. The variability nomenclature is denoted as follows: (P)ose, (E)xpression, (I)llumination, (G)lasses, (O)cclusion, (D)istance, (M)akeup and (T)ime-lapse.

Spectra	Name	Subjects	Images	Variability
NIR-VIS	CASIA HFB [93]	100	5097	Neutral
	CASIA NIR-VIS 2.0 [94]	725	17580	P,E,G,D
	Oulu-CASIA NIR-VIS [152]	80	7680	E,I
	BUAA-VisNir [59]	150	2700	P,E
	LAMP-HQ [146]	573	73616	P,I
	LDHF-DB [97]	100	-	D
SWIR-VIS	PRE-TINDERS [131]	48	576	E
	TINDERS [131]	48	1255	E,D
MWIR-VIS	WSRI [26]	64	3230	E
	MILAB-VTF(B) [112]	400	-	P,D
LWIR-VIS	UND X1 [108]	241	4584	E,I,T
	ARL-MMFD [52]	60	2280	E,D
	TUFTS [109]	113	900	P,E
	VIS-TH [98]	50	2100	P,E,I,O
	SF [2]	142	7668	P,E
	ARL-VTF [114]	395	549712	P,E,G
NIR-SWIR -LWIR-VIS	BYDB (ours)	412	1476292	P,E,G,I,O,D,M,T

preliminary evaluation aims to highlight the suitability of some multi-spectral thermal-visible paired face databases used throughout the dissertation.

Considering paired visible and long-wave infrared facial images, datasets introduced in Section 3.1.4 are used for this experimentation and Table 3.6 extends the previous Table 3.1 by adding other characteristics, overall summarized in Table 3.5. In particular, ARL-VTF and BYBD datasets consist of the largest collections of images with distinct subject identities, both especially acquired in a time-synchronized manner. Hence, the performances achieved should represent a significant value. As a standard practice, established protocol from

authors are followed for comparison purposes, with no overlapping between the training and testing set. The visible and thermal face images are processed and aligned following the methodology introduced in Section 4.5.3. Resulting images are then scaled to 128 × 128, and the evaluation performs an N:N comparison between visible-thermal spectra.

TABLE 3.6: Description of the paired visible-thermal face database used throughout the dissertation.

Name	**Resolution**	**Distance (m)**
ARL-MMFD [52]	640x480	2.5, 5, 7.5
VIS-TH [98]	160x120	1
TUFTS [109]	336x256	1.5
SF [2]	464x348	2
ARL-VTF [114]	640x512	2.1
BYDB	640x480	2,5,7

The direct comparison approach, in which a thermal probe is compared directly to the visible gallery, serves as a benchmark to deepen better understanding of the performances from different facial recognition methods *w.r.t.* robustness on different variability. In this approach, the thermal aligned face images and visible face images are fed to ArcFace [29] (a facial feature extractor introduced in Section 2.3.3), which extracts facial features and create a deep feature representation that is saved as template. Then, the similarity score between two templates is calculated using the cosine distance, to provide the verification score.

TABLE 3.7: Face verification performances reported when a direct comparison is performed between a thermal probe and a visible face image. AUC and EER metrics are computed using the ArcFace matcher.

Direct thermal to visible matching		
	AUC%	**EER**%
ARL-MMFD [52]	73.71	32.73
VIS-TH [98]	60.83	39.88
TUFTS [109]	57.28	44.68
SF [2]	56.57	45.19
ARL-VTF [114]	54.80	46.31
BYDB	50.03	45.24

Table 3.7 reports biometrics performances *w.r.t.* the Area Under the Curve (AUC) and the Equal Error Rate (EER) computed by using ArcFace matcher. The experimentation has shown that a direct comparison between thermal probe against a visible gallery results in degraded biometrics performances. This is due to the fact that traditional facial recognition systems, like ArcFace, are not trained to extract discriminative features on images acquired beyond the visible spectrum. As a recall, the accuracy of ArcFace (based on the network ResNet-100) achieves 99.83% on the visible face database MS-Celeb-1 M [136]. However,

implementing a thermal face feature extractor making effective direct visible-thermal face matching is impractical. The challenge of thermal-to-visible FR lies in the limited availability of paired thermal-visible face images for training robust facial feature extractors. Additionally to this challenge, thermal images contain poor details and low textured facial information, resulting in less high-frequency information which impede learning of pertinent facial features. Their degraded quality make it challenging to provide essential biometric information associated to the face in which deep models (feature extractors) would be able to learn. This dissertation aims to tackle above challenges by implementing generative models in order to reduce, (i) the modality gap, as well as (ii) to leverage any existing off-the-shelf visible-spectrum FR algorithms.

3.4 Summary

In this Chapter we provided an overview of datasets that are available for CFR research for the four sub-spectra of the infrared modality. We highlighted benefits and limitations of prior datasets *w.r.t.* practical CFR scenarios, thereby serving as motivation to design BYDB, an industrial in-house database referred to as a large-scale time synchronized multi-spectral face dataset. Particular efforts in the dataset collection had to do with addressing the need for images beneficial for CFR that incorporate an *unconstrained* setting, not present in existing datasets.

Chapter 4

Face and Landmarks Detector

Face and landmark detection are key modules to any FR system as they enable the localization of the face in the image and align it, prior to the feature extraction and matching steps. Consequently, developing face and landmark detectors adapted to thermal face images is an essential step towards creating an end-to-end CFR system. This is particularly challenging due to the inherent poor texture, low resolution and low contrast of thermal images, especially in an unconstrained environment.

In this Chapter, we propose a novel *thermal face and landmark detector* (TFLD) method streamlined to be robust to adversarial conditions. Unlike previous approaches, TFLD focuses on texture analysis with predicting faces or facial landmarks as regions of interest instead of specific marks. Thus, we here focus on textured areas rather than semantic points, which is instrumental in accurately detecting points, and by selecting relevant key-points, TFLD-based face alignment enhances biometric performances. Section 4.1 introduces the motivation behind the research presented in this work, while Section 4.2 revisits recent approaches and techniques addressed to perform facial landmarks detection in thermal imaging. Section 4.3 reports datasets used for the study and extends the methodology explored in the previous Section 3.2.3 to enhance current limitations on scarcity of annotations. Section 4.4 introduces the framework of the proposed TFLD. Finally, experimental results pertaining to face and landmark detection, as well as the impact of TFLD-based face alignment on CFR are presented in Section 4.5.

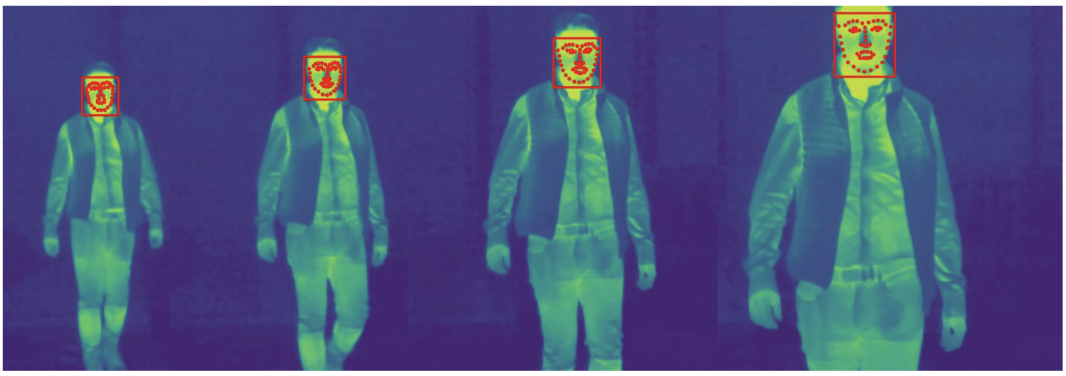


FIGURE 4.1: Monitoring system with thermal sensor. TFLD method applied on video sequence captured in the wild. A person, approximately 14m away, walks towards the camera while TFLD is tracking face and landmarks.

4.1 Introduction

Thermal cameras play a key role in many applications, ranging from surveillance to public safety. Figure 4.1 illustrates such assistance for instance in thermal monitoring systems. As introduced in this Chapter, the accurate detection of the face and its facial landmarks is essential information for subsequent stages of FR pipeline, that we proceed to design for our CFR system. This includes localizing an individual with the associated face in the thermal image, as well as aligning that face for the next stage of facial feature extraction.

Although landmark detection algorithms have made significant progress in the realm of visible spectrum, it remains a challenge in the context of thermal images due to the inherent factors such as *low contrast* and *poor resolution*, as well as the *lack of texture information*. In particular, thermal images contain less high-frequency information and associated degraded quality renders semantic definitions for certain landmarks inaccurate. Other challenges in developing new methods rely on the scarcity of thermal dataset equipped of facial annotations. This is why the multi-spectral dataset, BYDB, described in Chapter 3 is particularly instrumental useful for this purpose.

Motivated by the above, we present a novel *thermal face and landmark detector* (TFLD), especially designed to be robust under unconstrained circumstances such as *pose, expression, occlusion, poor image quality* and *long-range distance*. To overcome challenges relative to the scarcity of high-frequency information encountered by neural networks in the context of thermal images, paradigm of face and facial landmark detection are shifted from traditional methods to an object detection task. Therefore, YOLOv5 has been explored for its strong capacity in object detection, which further strengthens our approach where TFLD is designed to detect facial landmarks as object and considers them as the center of a textured area instead of specific points. TFLD represents a significant difference from traditional methods and offers the ability to detect accurately a large amount of thermal facial landmarks. Moreover, TFLD provides facial key points for the purpose of face alignment, which also

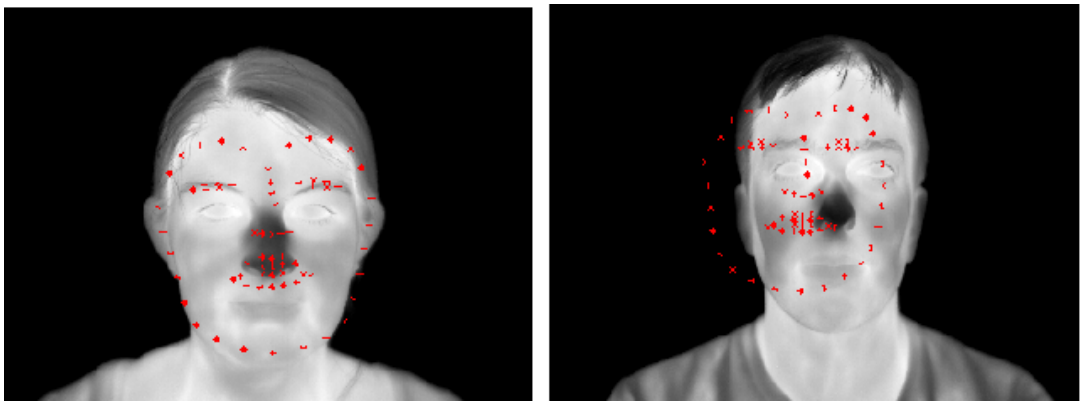


FIGURE 4.2: The results of directly applying on thermal face images a facial landmark detector trained for the visible spectrum. Landmarks predictions show significant deviations from the actual positions of landmarks. Source from experimentation [25].

demonstrates a positive impact on the FR scores. Therefore, rendering TFLD an accurate automatic annotation tool for CFR.

The main contributions of this work include the following.

- We propose a novel framework incorporating two successive YOLOv5-object detectors for face and landmark detection in the thermal spectrum, placing emphasis on robustness in unconstrained environments. TFLD predicts landmarks as regions of interest instead of specific marks, where textured areas are a principal concern rather than semantic points. We incorporate a thermal face restoration module as a pre-processing filter, allowing for a significant detection improvement.

- We (a) achieve at least state-of-the-art performance on three benchmark thermal face datasets with respect to *landmark localisation* and (b) improve automatic FR *matching scores* when using TFLD for pre-processing.

4.2 Background

While landmarks detection has become reasonably reliable in the context of the visible spectrum, it remains a challenge to generalize these trained algorithms onto thermal images due to the inter-spectral modality gap. The Figure 4.2 highlights the experimentation led by Chu and Liu [25]. It shows that landmark detector especially trained for image within the visible spectrum provides predicted landmarks deviating significantly from the ground truth.

As a consequence, number of approaches have been proposed to address the task of landmark detection in the thermal spectrum. Poster *et al.* [115] explored deep learning based approaches, including deep alignment network (DAN), multi-task convolutional neural network (MTCNN) and multi-class patch-based fully convolutional neural network (PBC). Chu and Liu [25] utilized the temperature information from face to design joint neural network model for both facial landmark detection and emotion recognition. Part of their proposed network

was based on a U-Net architecture including two output branches dedicated for the above respective tasks. Further extending deep learning based-model, Kuzdeuov *et al.* [86] compared the classical machine learning model Dlib based on a set of regression trees with a pyramidal like U-Net based model.

All previous approaches have been initially designed as well as extensively applied to detect facial landmark on visible images. However, when adapted and retrained for thermal images, they were facing the large inter-spectral modality gap, thus inferring room for improvement in terms of detection performances. To bridge the modality gap and tackle the lack of annotated thermal database, Mallat *et al.* [99] relied on generative adversarial models (GANs) to convert existing visible annotated facial databases into the thermal spectrum. Such strategy allows to build synthetic databases along with large amount a facial annotations. Active appearance model (AAMs) and DAN were then trained using the synthetic thermal samples along with the transferred facial annotations. Poster *et al.* [116] also utilized visible data by designing a coupled neural network with visible-to-thermal parameter transfer learning, however, their method was limited to be trained with only fully annotated paired-synchronized visible-thermal facial datasets.

4.3 Thermal face databases

In this Section, we delve into the databases that are used for training and testing TFLD method. Additionally, a strategy to allow enrichment of current limited facial annotations of databases is presented. Finally, to ensure the demonstration of being in line with building a comprehensive biometric system adaptive to adverse conditions, we propose a data augmentation suitable to diverse simulations of real world applications.

TABLE 4.1: Characteristics of the datasets used for the experiments of TFLD. Further information could be found in Table 3.6

Dataset	Spectrum	Facial landmark annotations
SF-TL54 [86]	LWIR/VIS	54 landmarks on thermal images
RWTH-Aachen [83]	LWIR	68 landmarks on thermal images
ARL-VTF [114]	LWIR/VIS	5 landmarks on thermal images
BYDB (ours)	LWIR/VIS	5 landmarks on thermal images

4.3.1 Facial databases endowed with rich ground truth landmarks annotations

To demonstrate the effectiveness of our proposed approach, we conduct experiments on various datasets that offer richer ground truth facial annotations and include thermal face images with different variations, ranging from frontal faces to extreme variations in poses, and expressions.

The *Speaking Faces - Thermal Landmark 54* [86] (SF-TL54) dataset presents simultaneously captured visible and thermal face samples. The dataset comprises 142 identities with $2,556$ images having a spatial resolution of 464×348,

where the train-test protocol is split as follows: 100 subjects are chosen for training and the remaining for testing. The database offers rich facial variations, including extreme poses captured from 9 different angles, as well as facial movements resulting from lips motions. Subjects are further wearing eyeglasses. Finally, SF-TL54 provides 54 facial landmarks particularly selected with semantically meaningful definition for thermal images. These points are partially based on a subset of the 68 landmarks from Dlib [80]. Throughout the experimentation, this database is denoted as *Pose sequences*.

The *RWTH-Aachen* [83] dataset is comprising the highest quality thermal face images and offers a spatial resolution of 1024×768. This dataset has been originally extended to include thermal face images showing posed expression, including *angry, contempt, disgust, fear, happy, sad, surprise*, and *neural*. It provides a fully manual annotation of 68 facial landmarks. With a total of 90 subjects over $2,935$ images, 60 subjects are dedicated for training and the rest for testing. RWTH-Aachen is thus denoted as *Expression sequence*.

The SF-TL54 and RWTH-Aachen datasets benefit from having comprehensive and rich facial landmark annotations, which makes them ideal for training TFLD under *pose* and *expression* variations, respectively. However, to further showcase the versatility of TFLD in detecting faces and landmarks under unconstrained circumstance, the experimentation also includes the largest collection of paired visible-thermal face image, namely ARL-VTF dataset. The next Section will outline the approach used to address the challenge posed by limited annotated images.

4.3.2 Missing ground truth extrapolation with visible-to-thermal landmark transfer

While the previous Section introduced databases with already annotated images, this Section aims at presenting a methodology to address annotations scarcity. Table 4.1 compares the databases with regard to their amount of provided facial landmarks annotations. Beyond the presentation of databases, many thermal face datasets are limited in their annotations, often only providing few labels for specific area such as the left and right center of eyes, nose, as well as left and right corner of the mouth. The lack of annotations results in very limited work focusing on thermal facial landmark detection. Therefore, following the objective of being adaptive to a wide range of variations and providing results for further comparison purposes, enhancement of annotations is elaborated.

By considering *synchronized visible-thermal* paired face databases along with a post-alignment processing, these allow extrapolating ground truth annotations from the visible spectrum to the thermal counterpart. Section 3.2.3 described the strategy relying on *homography* transformation for the post-alignment processing, in which multi-spectral faces are canonically aligned, enabling then the transfer of facial annotations from the visible spectrum to the invisible spectrum. This is illustrated in Figure 4.3, providing a full facial landmark of 68

annotations emerging from Dlib [80] to the images in the thermal spectrum. It is worth noting that, due to homography transformations in which all faces across spectra are overlapped, manual verification and correction are not required.

ARL-VTF [114], VIS-TH [98] and BYDB datasets were designed based on the same acquisition model where faces are simultaneously captured by thermal and visible camera. The method of extrapolating annotations is therefore applicable.

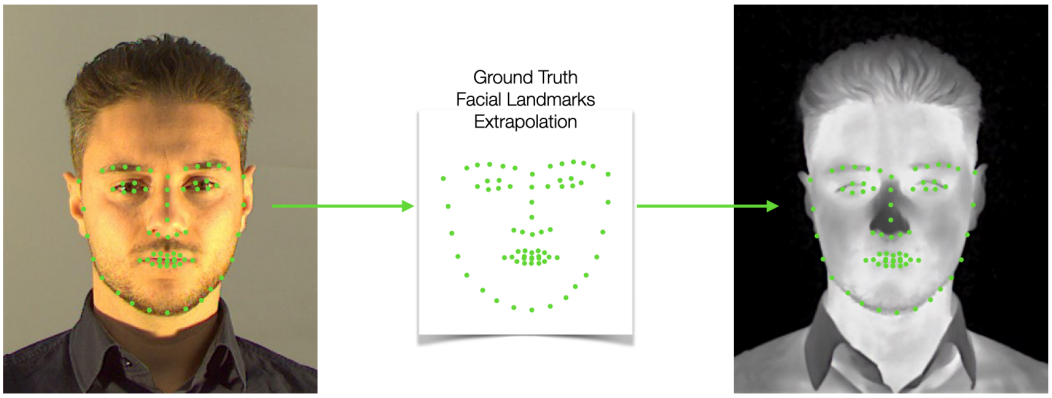


FIGURE 4.3: Given a synchronized *visible-thermal* paired face and a post-alignment processing, facial landmarks are extracted from the visible face (left) and transferred to the thermal counterpart face (right). The resulting ground truth annotations is provided by Dlib [80] where 68 facial landmarks are detected in the visible face.

4.3.3 Image augmentation

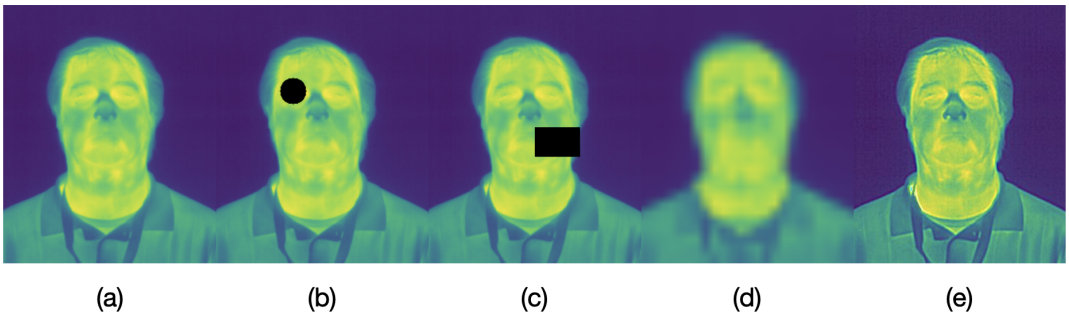


FIGURE 4.4: Data augmentation strategy. An original image (a) augmented by introducing (b) circular occlusion, (c) rectangular occlusion, (d) low resolution degradation, and (e) thermal face restoration processing. Samples coming from the ARL-VTF [114] dataset.

Data augmentation strategy plays a key role in the adaptability of TFLD in its capacity to operate in a wide range of situations. Targeting practical applications, detection through different facial occlusions or even at a long-range

distance is relevant. However, used databases do not contain such resources that would allow it to train TFLD directly. Therefore, to gain robustness in unconstrained environments, and hence to adapt better to in-the-wild scenario, datasets are augmented with following set of simulations. *Occlusions* are superimposed on the raw images by random circles or rectangles located in the eyes or mouth facial regions. On the other hand, *low resolution* degradation simulates long-range distance acquisition variations with down-sampling and up-sampling operations. Figure 4.4 (a-d) highlights the overall image augmentation.

Thermal face restoration

Another processing operation is considered as a *thermal face restoration* (TFR), depicted in Figure Figure 4.4 (e). Motivated by the challenges posed by thermal sensors such as poor image quality and low spatial resolution, accurate facial landmark localization requires an improvement in image quality. In particular, *image quality* in the context of thermal face images encompasses factors related to resolution, thermal contrast, signal-to-noise ratio, temperature accuracy, overall clarity, and consistency of thermal data. Therefore, *quality* is defined as the extent to which the thermal image faithfully represents the heat radiation patterns, characteristics, and features emitted by a subject's face. TFR allows enhanced visual details, contrast, and sharpness [9], which are meaningful features from the face for achieving accurate facial landmarks detection in thermal face images.

Given a thermal (thm) image $I^{thm}_{w\times h}$ where $w, h \in \mathbb{N}^*$ denote the width and height, respectively, the TFR filter is based on a combination of several difference of Gaussians (DoG) filters. The DoG filter as main operation, noted $\Gamma_{\sigma_i,\sigma_j}$, serves as a spatial band-pass filter and involves the subtraction of one Gaussian blurred G_{σ_i} version of an original image from another G_{σ_j} less blurred version of the original (see Equation (4.3)). The Gaussian blurred image B_σ is obtained by convoluting the original $I^{thm}_{w\times h}$ image with a Gaussian kernel G_σ having a standard deviation σ. This can be expressed as

$$B_\sigma = I^{thm}_{w\times h} * G_\sigma, \tag{4.1}$$

where the Gaussian kernel is from a two-dimensional Gaussian distribution

$$G_\sigma(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}. \tag{4.2}$$

In particular, the DoG filter is obtained by performing the subtraction of two images coonvolved with Gaussian kernels.

$$\Gamma_{\sigma_i,\sigma_j} = B_{\sigma_j} - B_{\sigma_i}, \tag{4.3}$$

with $\sigma_i < \sigma_j$. Finally, the enhanced image with TFR filter is obtained via a combination of two DoG filters consisting of both, different Gaussian kernel size

and standard variation values, and can be formalized as

$$TFR(I^{thm}_{w\times h}) = \Gamma_{\sigma_1,\sigma_2} + \Gamma_{\sigma_3,\sigma_4}. \tag{4.4}$$

Impact of the TFR filter on thermal face image is showed in Figure 4.4 where (a) and (e) highlight quality enhancement, as contributed by TFR.

4.4 TFLD

4.4.1 Subsequent face and landmarks detection

Monitoring systems as depicted in Figure 4.1 require continual observation in routine recording manner during which a multitude of people can appear. Hence, detection in two-stage process, namely *face* and *landmarks*, is the adopted methodology and allows for reliable multi-face detection and landmark detection in unconstrained settings. Benefits from this subsequent approach prevents the detection of false landmarks where no face appears. This incremental strategy has been further exploited by Keong *et al.* [77] in which the overall model is a stack of two subsequent models, an auxiliary model for detecting face boundary, and a main model for detecting facial landmarks, respectively.

4.4.2 Face and landmark designed as objects with biometrics meaning

The paradigm of face and facial landmark detection are shifted from traditional methods to an object detection task. It is essential to relate biometric meaning to object face as well as object landmark face. A face F is therefore considered as a bounding box, where the area is expressed by the inter-eye distance d_{IED} (center of the left and right eyes). F contains at least the left and right eyes, nose and mouth. To be specific, we apply the standard as highlighted in Figure 4.5 (a) where the upper left coordinates (up) and the bottom right coordinates (down) of F are based on d_{IED} with fixed $\alpha > 1$, defined as

$$(x_{up}, y_{up}) = (x_{\text{left eye}} - \frac{\alpha d_{IED}}{2}, y_{\text{left eye}} - \frac{\alpha d_{IED}}{3}), \tag{4.5}$$

$$(x_{down}, y_{down}) = (x_{\text{right eye}} + \frac{\alpha d_{IED}}{2}, y_{\text{right eye}} + \frac{2\alpha d_{IED}}{3}). \tag{4.6}$$

Nevertheless, when designing face detector for surveillance, it is important to consider that individuals may not always be facing the camera, thus providing different poses in profiles where both eyes are not visible simultaneously (see Figure 4.5 (b)). In this context, the above face box definition is not applicable. Therefore, we aim to provide a specific definition to improve the robustness of our thermal face detector in real-world scenario. Jointly considering the center of eye and associated mouse corner, we define the distance between them as d_{EM}. Hence, from the same face-side we obtain the upper left coordinates (up)

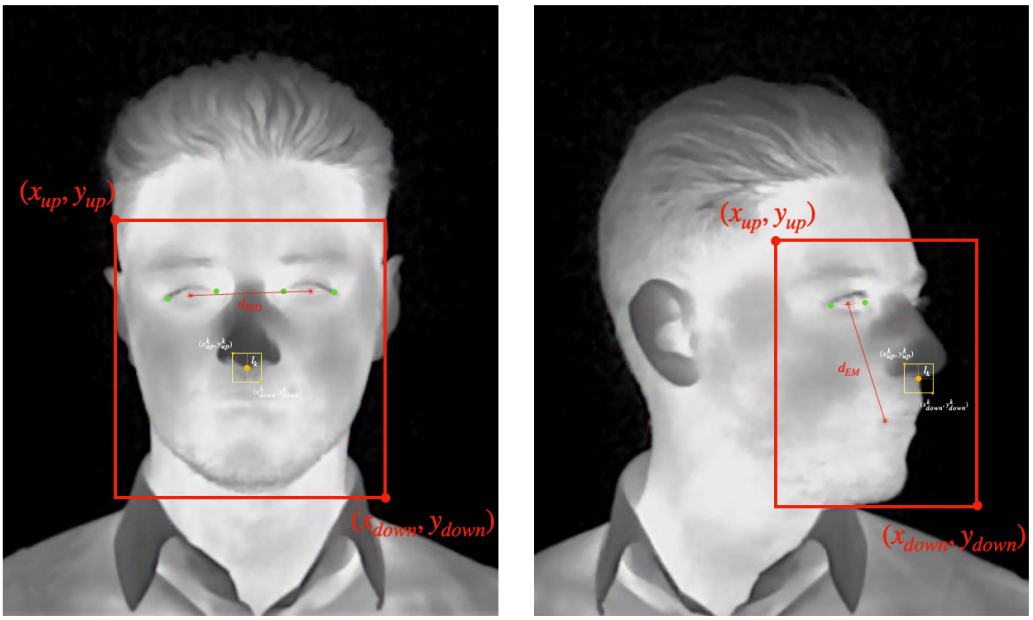


FIGURE 4.5: Semantic definition of a thermal face and thermal facial landmark based on different distances. (a) shows a frontal face and allows to consider the inter-eyes distance d_{IED}, while (b) depicts a face in profile, where one eye is occluded. In that case, another distance is considered, based on the visible eye and corner of the mouth d_{EM}.

and the bottom right coordinates (down), with fixed $\alpha > 1$, as follows

$$(x_{up}, y_{up}) = (x_{\text{eye}} \pm \frac{\alpha d_{EM}}{2}, y_{\text{eye}} \pm \frac{\alpha d_{EM}}{3}), \tag{4.7}$$

$$(x_{down}, y_{down}) = (x_{\text{mouth}} \mp \frac{\alpha d_{EM}}{2}, y_{\text{mouth}} \mp \frac{2\alpha d_{EM}}{3}), \tag{4.8}$$

where the sign is fixed *w.r.t.* the facial profile which is not hidden. This adaptive strategy will allow to detect effectively faces in various poses and orientations.

Regarding facial landmarks, we define them as a region of interest through a custom box, where the center represents the desired landmark l_k with proportional width and height shape, scaled to d_{IED} or d_{EM}. Consequently, for all $k \in [0, K]$ the associated landmark l_k is expressed by the shape of the custom box with

$$l_k = (\frac{x^k_{up} + x^k_{down}}{2}, \frac{y^k_{up} + y^k_{down}}{2}), \tag{4.9}$$

where K is the total number of facial landmarks, (x^k_{up}, y^k_{up}) and (x^k_{down}, y^k_{down}) denote the upper left and bottom right coordinates corner of the k-th custom box, respectively.

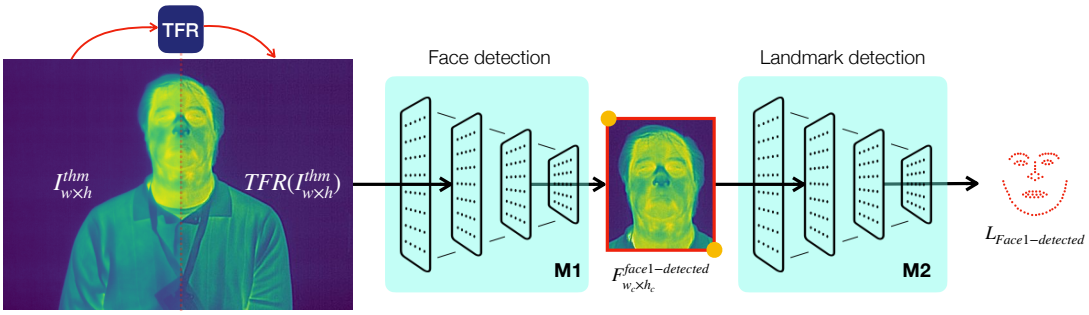


FIGURE 4.6: Illustration of the TFLD pipeline. A TFR filter is first applied to a thermal image $I^{thm}_{w\times h}$. Hence, the network is fed by an enhanced $TFR(I^{thm}_{w\times h})$ thermal image, where $M1$ is responsible of the face detection $F^{face1-detected}_{w_c\times h_c}$, whereas $M2$ is dedicated to extract a set of facial landmarks $L_{face1-detected}$.

4.4.3 Baseline model

We design a model, where a TFR pre-processing filter is succeeded by a series of two "You Only Look Once" (YOLOv5) models. Both of them are based on the medium backbone of the fifth version, and are denoted as $M1$ and $M2$, respectively. The former is responsible of detecting region of interests (ROIs) that contains faces from the background, while the later aims at extracting a set of landmarks pertaining to the prior cropped region. Models are optimized by three objectives functions that include (i) the mean square error as *bounding box regression loss*, (ii) the binary cross entropy as *objectness loss* and (iii) the cross entropy as *classification loss*.

The overall architecture of TFLD is illustrated in Figure 4.6. Given a set of N thermal images we denote the n-th image as $I^{thm}_{w\times h,n}$, then for all $n \in [1, N]$, the TFR filter is first applied as a pre-processing step. Hence, the network is fed by the enhanced version $TFR(I^{thm}_{w\times h,n})$, where $M1$ is responsible of detecting faces and outputting the associated set of F_n faces on that n-th image. Note that, if no face is detected, $M1$ returns an empty set and no landmarks will be subsequently detected. This is formalized as follows.

$$M1(TFR(I^{thm}_{w\times h,n})) = \begin{cases} \emptyset & \text{if } F_n = 0 \\ \bigcup_{f=1}^{F_n} F^f_{w_c\times h_c,n} & \text{otherwise,} \end{cases} \tag{4.10}$$

where $f \in [1, F_n]$ and $F^f_{w_c\times h_c,n}$ is the f-th cropped face within the n-th image. $F^f_{w_c\times h_c,n}$ encompasses the points $\{(x^f_{up}, y^f_{up}), (x^f_{down}, y^f_{down})\}$, denoting the upper left and bottom right coordinates corner of the f-th face bounding box, respectively, further marked by the plotted red points on the Figure 4.5.

Once the face is detected, meaning that $F_n \neq 0$, the second model $M2$ produces the final landmark output. Therefore, considering all faces $f \in [1, F_n]$ in that image, $M2$ provide a set $L_{f,n}$ of K landmarks corresponding to the f-th face of the n-th image. This is formalized as follows.

$$M2(F^f_{w_c\times h_c,n}) = L_{f,n}. \tag{4.11}$$

In particular,

$$L_{f,n} = \bigcup_{k=1}^{K} l_k^{f,n}, \tag{4.12}$$

where $l_k^{f,n}$ refers to the coordinates $(x_k^{f,n}, y_k^{f,n})$ of the k-th landmark present in the f-th face of the n-th image. If a particular landmark $l_k^{f,n}$ is undetected, $M2$ returns an empty point $\emptyset$, which is then interpolated with other based predicted landmarks.

Finally, for all $f \in [1, F_n]$, $L_{f,n}$ is reported on the original $I_{w\times h,n}^{thm}$ thermal input image, providing therefore the final landmark locations.

4.5 Experiments

4.5.1 Evaluation Metrics

This Section presents metrics used to assess qualitative capabilities of detection *w.r.t.* (i) thermal face detection as well as (ii) thermal facial landmark detection.

Face detection - Model $M1$

Although mean Average Precision (mAP) is usually used to evaluate any object detection algorithm, including face detection task, we selected other standards. Related work did not investigate face detection for the used datasets, we thus emphasized the analysis on the landmark detection instead. We selected a metric suitable for the TFLD pipeline, useful for the next landmark stage detection, where a prior face detected is considered as *well detected* if it contains salient regions of the face.

Face detection performances is evaluated by the *Detection Rate* (DR) metric. Given a thermal images $I_{w\times h,n}^{thm}$ containing F_n faces, the number of faces properly detected by the model $M1$ is expressed in terms of cardinality defined as

$$\mid M1(I_{w\times h,n}^{thm}) \mid = F. \tag{4.13}$$

According to Equation (4.10), F denotes the number of total faces well detected on the n-th thermal image. In particular, a detected face $F_{w_c\times h_c,n}^{f}$ is validated if and only if the bounding area of size $w_c \times h_c$ comprises at least eyes and mouth ground truth annotations.

Finally, the face DR per image is the ratio F/F_n and the total face DR assessed by $M1$ is given as follows.

$$DR_{M1} = \frac{1}{N}\sum_{n=1}^{N} \frac{\mid M1(I_{w\times h,n}^{thm}) \mid}{F_n}, \tag{4.14}$$

where N is the total number of images tested, and F_n the total number of faces that are annotated in the n-th images.

Landmark detection - Model $M2$

The landmark localisation performance is evaluated by two joint metrics, namely the *Normalized Point-to-Point Error* (NPPE) and the *Normalized Mean Error* (NME). Given a detected face $F^{f}_{w_c \times h_c, n}$ from the f-th face present in the n-th images of a set of N testing thermal images $\{I^{thm}_{w \times h, n}\}^{N}_{n=1}$, the NPPE metric computed for a particular landmark $k \in [1, K]$ is referred to as $P^{f,n}_{k}$ and defined as follows.

$$P^{f,n}_{k} = \frac{\|l^{f,n}_{k} - \hat{l}^{f,n}_{k}\|}{d_{IOD}}, \tag{4.15}$$

where l is the ground truth coordinate and $\hat{l}$ the predicted coordinate provided by $M2$. For comparison purposes *w.r.t.* other methods we use the same normalization protocol, where d_{IOD} expresses the inter-ocular distance (considering the outer corners of the eyes). The NME metric is further used to assess the average performances on the set, and is obtained by

$$NME = \sum_{f,k}^{\mathcal{F},K} \frac{P^{f,n}_{k}}{\mathcal{F} \times K}, \tag{4.16}$$

where $\mathcal{F}$ indicates the total number of faces in the set.

Finally, TFLD aims to provide a higher DR obtained in Equation (4.14) while a lower NME expressed in Equation (4.16).

4.5.2 Evaluation of TFLD

To evaluate the effectiveness of the TFLD approaches, we conduct a series of tests on the ARL-VTF, SF-TL54 and RWTH-Aachen datasets including *baseline, expression* and *pose* sequences under several variations: Raw[1], Occlusion and Poor image quality along with (w/) or without (w/o) applying the TFR filter.

Face and Landmark detection

TABLE 4.2: Landmark detection performance represented by the Normalized Mean Error (NME), on ARL-VTF, SF-TL54 and RWTH-Aachen datasets.

TFLD - $M2$	**ARL-VTF**			**SF-TL54**	**RWTH-Aachen**
Landmark detection - NME	*Baseline*	*Expression*	*Pose*	*Pose*	*Expression*
Raw *w/o* TFR	0.05244	0.05468	0.06804	0.03082	0.03311
Raw *w/* TFR	**0.05201**	**0.05460**	**0.06737**	**0.03001**	**0.03289**
Occlusion *w/o* TFR	0.05958	0.07328	0.07811	0.04243	0.03975
Occlusion *w/* TFR	0.05549	0.06170	0.07789	0.03856	0.03897
Low Resolution *w/o* TFR	0.07099	0.08861	0.09378	0.05205	0.04167
Low Resolution *w/* TFR	0.06934	0.08605	0.09118	0.04999	0.04039

[1]Raw denotes the original image acquisition emerging from the thermal sensor.

TABLE 4.3: NME score comparison of TFLD with other approaches on different datasets.

Authors	Methods	SF-TL54	RWTH-Aachen
Mallat *et al.* [99]	AAM	-	0.143
	DAN	-	0.146
Chu *et al.* [25]	Dlib (adapted)	-	0.095
	U-net (multitask)	-	0.040
Kuzdeuov *et al.* [86]	Dlib (optimizer)	0.033	0.057
	U-net	0.035	0.058
Ours	TFLD	**0.03001**	**0.03289**

Model $M1$ plays a first role in the pipeline of detecting faces. When tested in all test sets, it demonstrated a perfect detection rate DR_{M1} since 100% faces were detected, with or without applying the TFR filter, even under occlusion or low-resolution variations. Model $M2$ is responsible for locating facial landmarks and has been trained separately on each dataset due to the differences in the amount and semantic definition of landmarks.

TFLD accurately detects facial key points on baseline, expression, and pose sequences, as shown in Figures 4.7, 4.8 and 4.9. In particular, the TFR filter enhances the sharpness of the images and improves the accuracy of TFLD's landmark detection, as seen in Figure 4.7. However, poor image quality and occlusion present challenges for landmark detection. Nevertheless, TFLD is able to correctly identify facial key points even in the presence of occlusion due to glasses (Figure 4.8).

Table 4.2 displays the quantitative results in terms of NME, which are stable across all datasets. Note that landmarks are predicted on images where prior faces are detected by $M1$. TFR consistently improves performance, but occlusion and low resolution only have a minor impact on performance.

Comparison with State-of-the-Art

Table 4.3 presents the NME scores of landmark detection performance on the SF-TL54 and RWTH-Aachen datasets, compared to the State-of-the-Art methods. It is worth noting that our study is the first to report results on the ARL-VTF dataset, as shown in Table 4.2 with TFR processing under the Raw setting. The results of TFLD are promising compared to the State-of-the-Art approaches, which are based on either machine learning or deep learning.

Unified model applied on real-wold thermal images

We proceed to demonstrate that TFLD is fully capable of operating on real-world images (see Figure 4.1), especially not seen during its training. To prove this, we are leveraging a joint training model on ARL-VTF, SF-TL54 and RWTH-Aachen datasets. This unified training induces to the model the name *Unified-TFLD*.

Unified-TFLD model aims to predict 5 facial key-points, namely eye centers, nose and mouth corners. Towards assessing TFLD's performances, we aim to show its capacities to various scenarios on BYDB dataset, ranging from

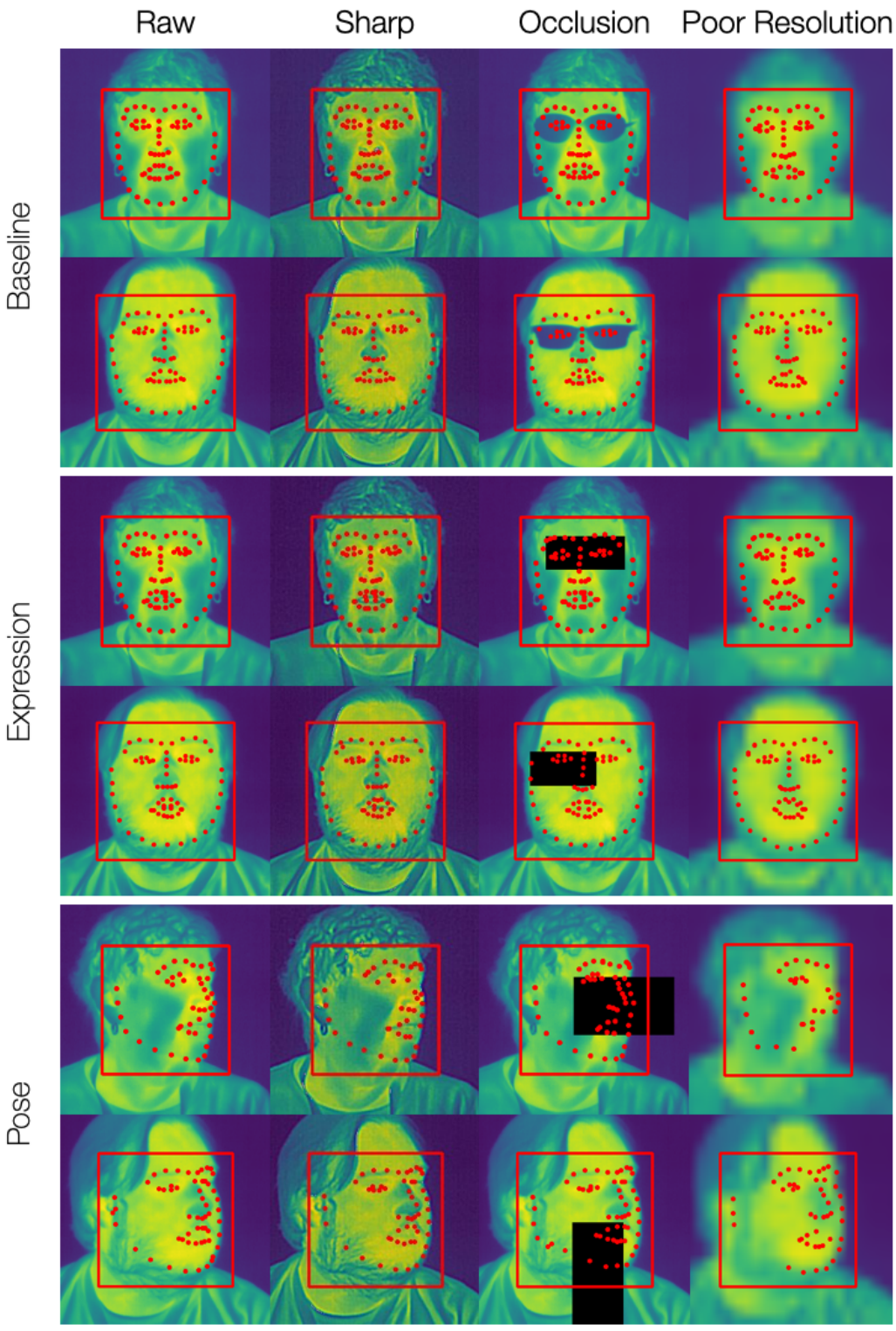


FIGURE 4.7: **Visualized faces and landmarks as detected by TFLD on the ARL-VTF [114] dataset**. The face is first detected (red box) followed by landmark detection (red points). TFLD is challenged with *Baseline*, *Expression* and *Pose* sequences, comparing *Raw*, *Sharp (TFR)*, *Occlusion* and *Poor resolution* degradations.

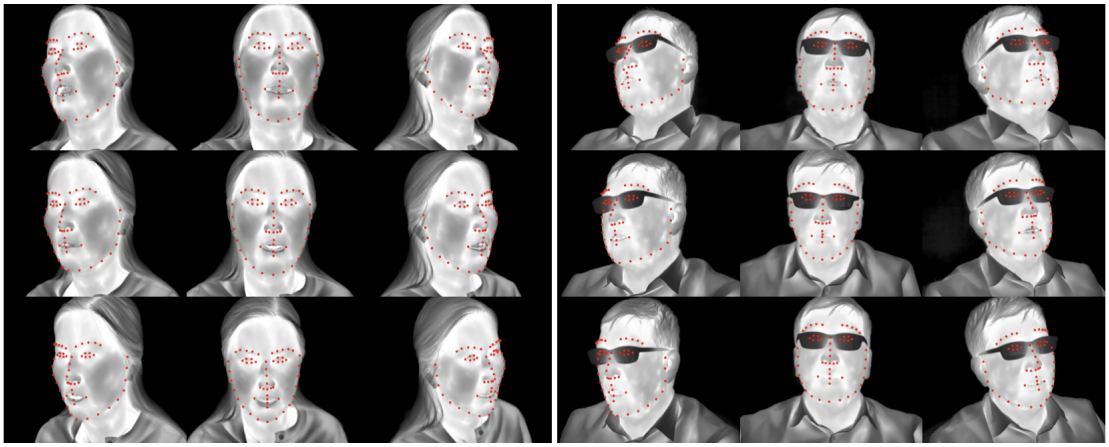


FIGURE 4.8: **Visualization of the landmark detection performed by TFLD model on the SF-TL54 [86] dataset**. TFLD appears robust to *Pose* variations and *Occlusion*.

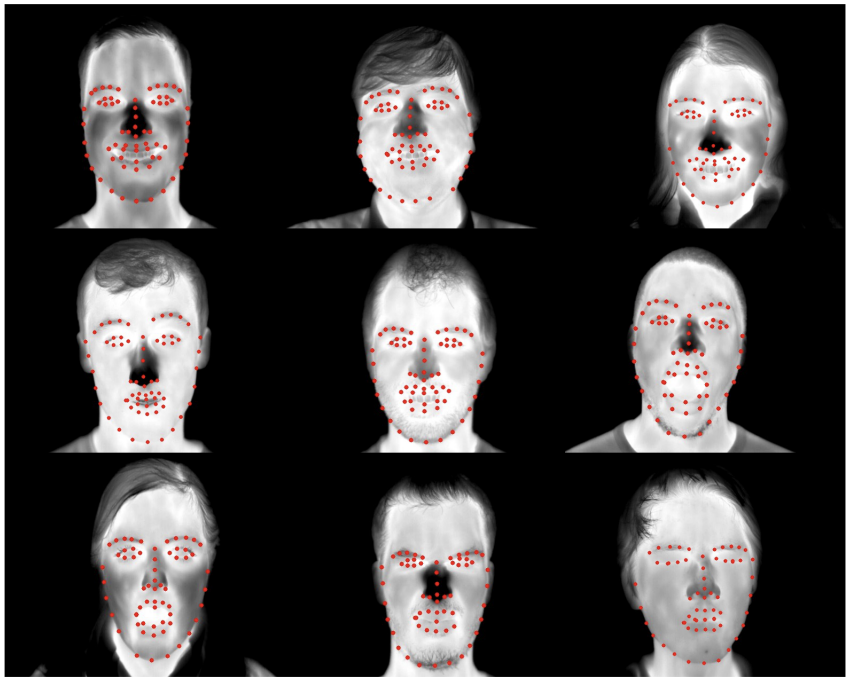


FIGURE 4.9: **Visualization of the landmark detection performed by TFLD model on the RWTH-Aachen [83] dataset**. TFLD appears robust to *Expression* variations.

frontal to profile poses, occlusions, and expressions. The Figure 4.10 offered a visualization under each settings. Key-points are accurately located, even under challenging conditions such as pose in extreme profile or occlusion. Particularly, facial expression or eyes closed do not perturb the good prediction.

FIGURE 4.10: Visualized landmarks as detected by Unified-TFLD on (ours) BYDB dataset. Unified-TFLD is challenged with practical scenarios, including *Baseline*, *Pose*, *Expression* and *Occlusion*. The five key-points are accurately localized, even under pose or occlusion, from facial expressions or eyes closed.

TABLE 4.4: Landmark detection performance represented by the Normalized Mean Error (NME), on BYDB dataset with *Unified-TFLD*.

Unified-TFLD Landmark detection	**BYDB**			
	Baseline	*Pose*	*Expression*	*Occlusion*
NME	0.02786	0.09484	0.02830	0.06316

Table 4.4 reports NME scores achieved by Unified-TFLD, from the same settings. Note that, only the eyes and mouth landmarks have been selected for the quantitative assessment, the nose landmark depicted another semantic definition, namely middle of nose instead of the base-center as learnt by TFLD. The results highlight the effectiveness of our proposed methodology and its practicality in real-world applications. Building upon the results of this experimentation, there is a potential for industrial application beyond the academic area. Unified-TFLD, as a preprocessing tool, allows for precise detection of faces and facial landmarks in thermal imaging, information that can be used to align an image in a standard manner and increase accuracy as well as efficiency of a biometric system.

4.5.3 Impact of face alignment on CFR

In this Section, we motivate the advantage of face detection and alignment as a preprocessing technique. This is of particular pertinence in designing an end-to-end biometric system robust to unconstrained environment. To this end, TFLD aims at demonstrating that, used as automated face alignment, it can improve CFR. We specifically investigate how different facial landmarks serving as based-alignment can affect the FR matching score.

We proceed to evaluate the impact of face alignment in thermal images *w.r.t.* CFR. We therefore compare the FR scores originating from method (i) cropping the facial images based on the detected bounding box and (ii) aligned facial images based on facial key points.

Cropping face images involves resizing the face bounding box detected by $M1$ to a specific size, while aligning the face based on key points involves applying affine transformations, including translation, scaling, and rotation, to align the face with the geometric centers of the eyes, nose, and mouth.

The matching process is carried out using the ArcFace [29] matcher. Table 4.5 reports the scores of facial recognition, showing the Area Under the Curve (AUC) metric, which is calculated between the visible gallery face and the aligned thermal probe face. A higher AUC value indicates better performance. It should be noted that RWTH-Aachen dataset was not included in this experiment because it does not have paired visible-thermal faces.

TABLE 4.5: Evaluation of the impact of face alignment in thermal images toward a cross-spectral face recognition system with respect to face recognition matching scores AUC % (higher is better).

Alignment based on	**ARL-VTF**	**SF-TL54**
Bounding box (no alignment)	52.86	60.89
GT annotations	56.03	69.20
Dlib (optimizer) [86] annotations	55.56	67.84
TFLD annotations	**56.93**	**69.59**

At first glance, the performance of face matching is poor when the images are provided without alignment. However, aligning the images based on the five key points mentioned above leads to an improvement in the scores. When comparing the AUC scores of facial images aligned with ground truth annotations, Dlib decreases the score, while our TFLD method slightly improves the score. As a result, TFLD is not only effective in a wide range of adverse conditions, but also demonstrates its potential as a robust and *accurate facial landmark annotator* that is crucial for CFR systems.

4.5.4 Discussion

The two-stage detection process, consisting of face and landmark detection, enables accurate detection of multiple faces and landmarks in uncontrolled environments. This approach prevents the detection of erroneous landmarks in

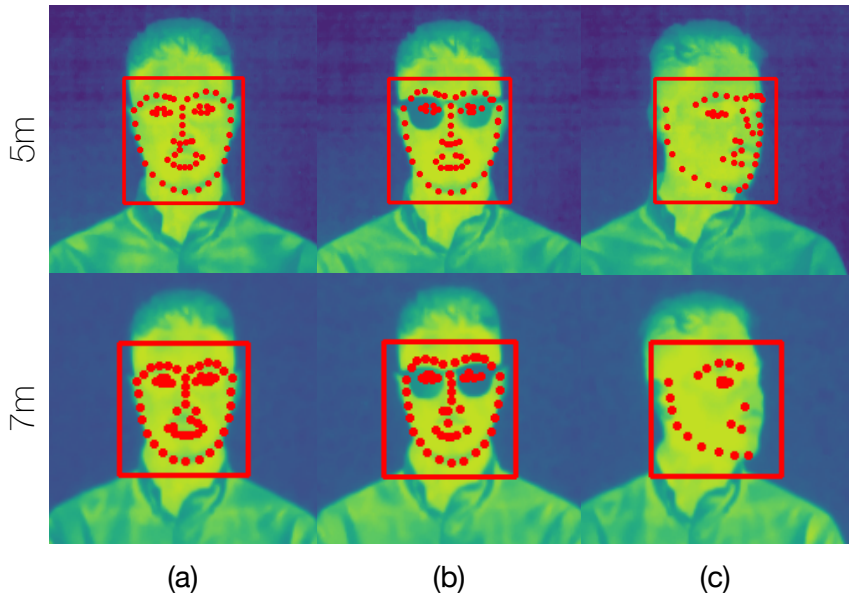


FIGURE 4.11: Examples of TFLD in unconstrained thermal images. The top row shows images acquired at an offset distance of 5m, whereas the bottom row at 7m including the covariates (a) frontal pose variation, (b) eye glasses, and (c) face in profile.

regions where no face is present. Figures 4.1, 4.11, and 4.12 depict a real-life scenario where individuals are captured at varying distances in both indoor and outdoor environments. In such situations, image quality significantly decreases, making it difficult to localize facial features. However, data augmentation resolves these issues and enables the TFLD model to detect faces and landmarks effectively, thanks to the TFR filter's enhancement.

Collecting facial annotations is both costly and laborious. The ARL-VTF dataset made it possible to transfer visible-to-thermal landmarks, which increased the number of landmarks used to train the TFLD method. Our pipeline takes advantage of synchronized images, which helps to overcome this problem (lack of facial annotations). The infusion of additional landmarks has a significant impact on improving *e.g.,* FR performances. However, at the time of writing this dissertation, few datasets meet these criteria of time-synchronized visible-thermal paired faces, which limits the possibility of creating a more unified method of landmark detection and considering them as benchmarks for face and landmark detection.

4.6 Summary

Accurate and reliable automatic face and landmark detection is a key preprocessing step in CFR. We have therefore proposed a novel thermal face and landmark detector (TFLD) that accurately localizes faces and landmarks in unconstrained settings.

The proposed TFLD sequentially detects face and landmarks, thereby improving the accuracy of landmarks localization. Experiments highlight that

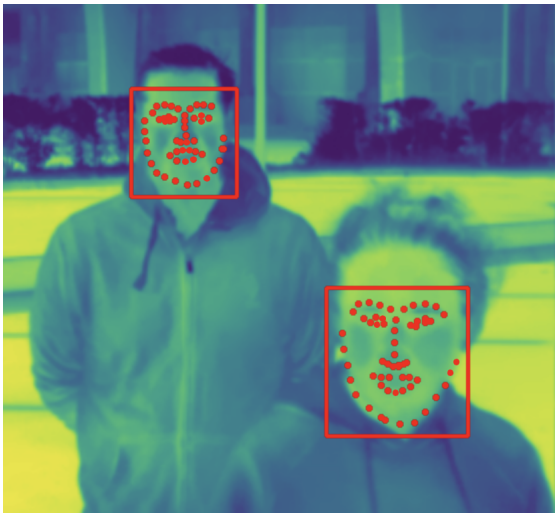


FIGURE 4.12: Examples of TFLD operating in an outdoor environment with sunny weather. Despite challenging atmospheric conditions, faces, eyes, eyesbrows, nose, mouth and jawline key points are successfully located by our model (image from TFW [87] dataset).

the proposed TFLD method achieves competitive results on three benchmarks, even under challenging pose and expression variations. In addition to academic experimentation led on three datasets, TFLD was applied on real unconstrained thermal images. Notably, the TFLD-unified version indicating its strong potential for real-world applications, as reported by its performance on the BYBD dataset. Finally, we have demonstrated the positive impact of face alignment based on TFLD on CFR.

Chapter 5

Identity Preserving Spectrum Translation

Facial biometric systems conventionally operate by comparing face images within the same spectrum, *e.g.,* the visible as RGB-imagery. The reliability of such systems is largely attributed to the vast number of visible face images available for training algorithms, enabling the learning of discriminative identity features. Recognizing individuals in the thermal spectrum entails enrolling their thermal face images in a gallery first. However, this is highly impractical as thermal cameras are not widely deployed for FR purposes. Therefore, CFR attempts to alleviate this limitation by comparing faces across spectra, thus eliminating the need to enroll faces beyond the visible spectrum. Nevertheless, the large domain gap between the thermal and visible spectrum makes CFR more challenging. Towards mitigating the modality gap, GANs are leveraged to translate a thermal face image into a realistic synthetic visible-like face image, employed as probe-image.

This Chapter presents a comprehensive approach for translating thermal face images to synthetic visible face images, while preserving the identity of depicted individual. Two novel *explainable* models are proposed with particular emphasis on explaining how identity features are learned during training. Furthermore, interest is placed on exploring specific loss functions, which are essential to ensure faithful identity preservation, as well as to generate realistic faces. Section 5.3 introduces a Latent-Guided Generative Adversarial Network (LG-GAN), which aims to explicitly decompose an input image into an *identity code* that is spectral-invariant and a *style code* that is spectral-dependent. In Section 5.4, a second model, namely Attention-Guided Generative Network (AG-GAN), was developed with the aim of using attention feature maps as guidance features synthesis process. LG-GAN and AG-GAN offer critical information in *finding representations* that allow for *extracting* robust biometric features beyond the visible, as well as interpreting the generation process from thermal to visible. These models improve the CFR matching accuracy and inherently support model explainability. Finally, extensive FR experiments have been conducted in Section 5.5.

5.1 Introduction

Towards enabling the comparison of heterogeneous facial samples, finding representations that allow for extracting pertinent biometric features, beyond the visible, is critical. The large discrepancies in-between spectra could be minimized by using GANs to synthesize visible faces from their thermal counterparts [147, 151, 137, 21, 34]. This strategy falls into the computer vision task of domain translation. However, to the best of our knowledge, there is no prior thermal-to-visible GAN that explicitly investigates salient *identity determining features* across the thermal and visible domains.

We therefore designed two explainable models that translate thermal face images into synthetic visible face images, while preserving the identity. In particular, the transfer of salient biometric data from thermal to visible-like face is of particular pertinence in ensuring efficient CFR. Here, the definition of *explainability* relies on the visualization of features which are learnt during the training. In order to explore discriminative features, we adopt an encoder-decoder structure. As stated in next formalization Section 2.1, CFR assumes that the identity space is shared by images of a given subject acquired in different spectra, whereas the spectrum space is not. Hence, *salient facial features* are disentangled from other confounding factors with (i) a Latent-Guided Generative Adversarial Network (LG-GAN), as well as (ii) an Attention-Guided Generative Network (AG-GAN). We here note that the latent space is defined as the feature space learned by a GAN.

LG-GAN decomposes a facial image into an *identity code* and a *style code*, where the identity code is learned to encode spectral invariant identity features across spectra. Then, to translate an image from a source spectrum to a target spectrum, the identity code is combined with a style code denoting the target domain. AG-GAN, on the other hand, learns attention modules to capture global interaction between facial context across spectra. A thermal face image is encoded as an attention feature map to deepen knowledge on salient cross-spectral facial features that are learned during spectrum translation. Particularly, attention maps serve to guide feature synthesis at focusing on specific regions relevant to CFR. In this context, attention maps are provided by attention feature weights, that can be learnt in supervised or unsupervised manner, giving rise to AG-GAN and AG-GAN+, respectively. To simplify, we note AG-GAN(+) for referring in general to the method. In addition, a commonality between LG-GAN and AG-GAN has to do with the use of specifically tailored loss functions that facilitate generation of *realistic faces*, as well as that ensure faithful *identity preservation*.

Finally, the identity code offers useful insights into explaining salient facial structures that are essential to the synthesis of high-fidelity spectrum face images, while the visualization of attention feature maps entails transparency and interpretability regarding salient facial features that are discriminative across spectra.

5.2 Background

The cross-spectral identity matches were implicitly enforced by minimizing the reconstruction error [21] known as L_1 loss or Euclidean loss, or by using an identity extraction network [147, 21, 34]. However, to the best of our knowledge, there is no prior thermal-to-visible generative model that explicitly investigates, which are the *identity determining features* across the thermal and visible domains.

In this Section, we discuss existing literature on conditional adversarial networks and explainable GANs in performing general image-to-image translation.

Conditional adversarial networks [68] have so far been the defacto model to solve image-to-image translation tasks in supervised settings. Prior works have involved notably Pix2Pix [68], aimed at learning to map a conditional input thermal image to an output visible image. The optimization step was further regularized by introducing additional constraints such as closed-set face recognition losses [151, 113] or face verification losses [21], in order to preserve the identity mapping. In comparison to these, some other recent works have focused on preserving the attribute mapping by using a pre-trained attribute prediction network [64, 34]. In addition to the preservation of identity and attribute mappings, some methods [63] focused on elaborate network architecture design, incorporating the self-attention module from the Transformers [35].

Explainable GANs aim to empower image translation by transparency and interpretability. Recent works predominantly focused on the visualization and understanding of internal representations [79]. Kim *et al.* [79] incorporated learnable attention modules into the generator and the discriminator for unsupervised image-to-image translation. Tang *et al.* [130] proposed an attention-guided generator to disentangle the semantic objects from the background via producing an attention mask and a content mask. The attention module was also integrated into the discriminator, focusing on attended regions only. Their proposed attention-guided generator and discriminator were used to solve unpaired image-to-image translation, which demonstrated promising results, in case that the geometric change between the source and target domain is minor.

5.3 LG-GAN

5.3.1 Baseline Model

We propose LG-GAN, a latent-guided generative adversarial network, designed for paired thermal-to-visible translation. Specifically, LG-GAN learns the *content* as well as the *style* pertaining to a face that we refer to as *identity* and *style* code, respectively, in the latent space. LG-GAN is inspired from MUNIT [62]. The overall architecture is illustrated in Figure 5.1. In LG-GAN, *identity* and *style* are essential in translating images from an input thermal domain to an output visible domain, thereby bridging the domain gap through (a) enforcing

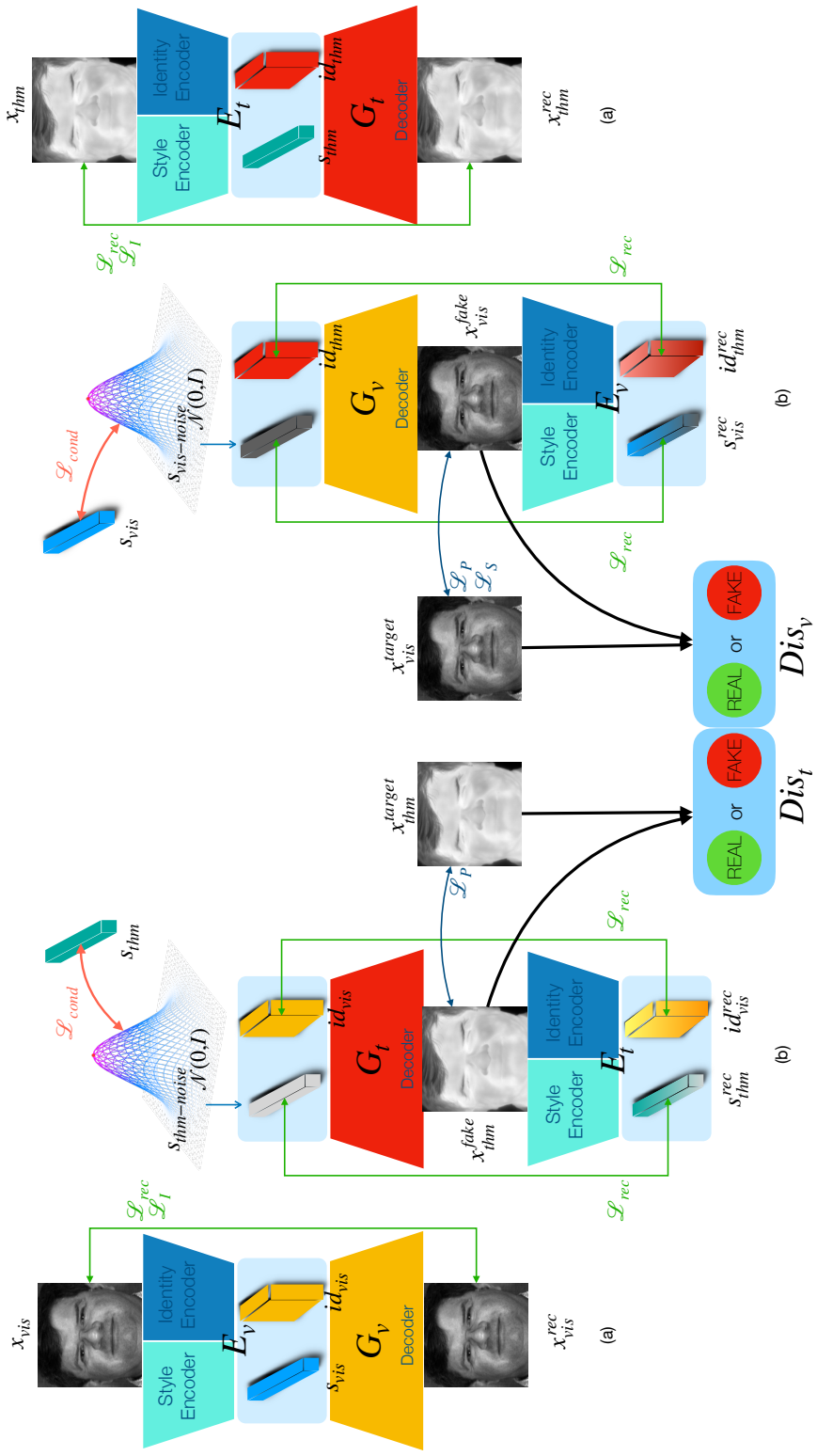


FIGURE 5.1: Flowchart depicting the training of the proposed LG-GAN framework. It consists of two auto-encoders (E_v, G_v) and (E_t, G_t) dedicated to the visible and thermal domains, respectively. The sub-network (a) aims to learn the image reconstruction, while (b) enforcing the latent space reconstruction.

both image-level and latent-level reconstructions, and (b) supervising thermal-to-visible image translation with an identity preserving loss function. Note that MUNIT [62] does not deal with the problem of CFR.

The generator comprises of three networks for each domain, viz., *Identity-Encoder, Style-Encoder* and *Decoder*, targeted to extract a domain-shared identity latent code and a spectrum-specific style latent code. The translated image is reconstructed by combining the identity code with the style code of the target spectrum. Figure 5.2 illustrates the auto-encoder architecture. In the discriminator, we adopt the multi-scale discriminator which enables generation of realistic images with refined details.

The main contributions of this work include the following.

- We propose a novel *supervised learning* framework for CFR that translates facial images from one spectrum to another, while preserving the identity.
- We introduce four loss functions that facilitate both image as well as latent reconstructions.
- We analyze the latent space, which is decomposed into a shared *identity* space and a spectrum-dependent *style* space, by visualizing the encoding using heatmaps.
- We evaluate the proposed framework on two benchmark multispectral face datasets and achieve promising results with respect to *visual quality*, as well as face recognition *matching scores*.

5.3.2 Formalization

Let $\mathcal{V}$ and $\mathcal{T}$ be the visible and thermal domains. Let $x_{vis} \in \mathcal{V}$ and $x_{thm} \in \mathcal{T}$ be drawn from the marginal distributions $x_{vis} \sim p_{\mathcal{V}}$ and $x_{thm} \sim p_{\mathcal{T}}$, respectively. Thermal-to-visible face recognition based on GAN-synthesis aims to estimate the conditional distribution $p_{\mathcal{V}|\mathcal{T}}(x_{vis}|x_{thm})$, where,

$$p_{\mathcal{V}|\mathcal{T}}(x_{vis}|x_{thm}) = \frac{p_{\mathcal{V},\mathcal{T}}(x_{vis}, x_{thm})}{p_{\mathcal{T}}(x_{thm})} \tag{5.1}$$

involves the joint distribution $p_{\mathcal{V},\mathcal{T}}(x_{vis}, x_{thm})$. As the joint distribution is not known, we adopt the assumption of "partially shared latent space" from MUNIT [62] as follows.

A pair $(x_{vis}, x_{thm}) \sim p_{\mathcal{V},\mathcal{T}}$ of images, corresponding to the same face from the joint distribution, can be generated through the support of

(a) the **identity latent code** $id \in \mathcal{I}$, which is shared by both domains (we also introduce the notation $id_{vis}, id_{thm} \in \mathcal{I}$ for better domain-identity formalization),

(b) the **style latent code** $s_m \in \mathcal{S}_{\mathcal{M}}$, where $(m, \mathcal{M}) \in \{(vis, \mathcal{V}), (thm, \mathcal{T})\}$, which is specific to the individual domain.

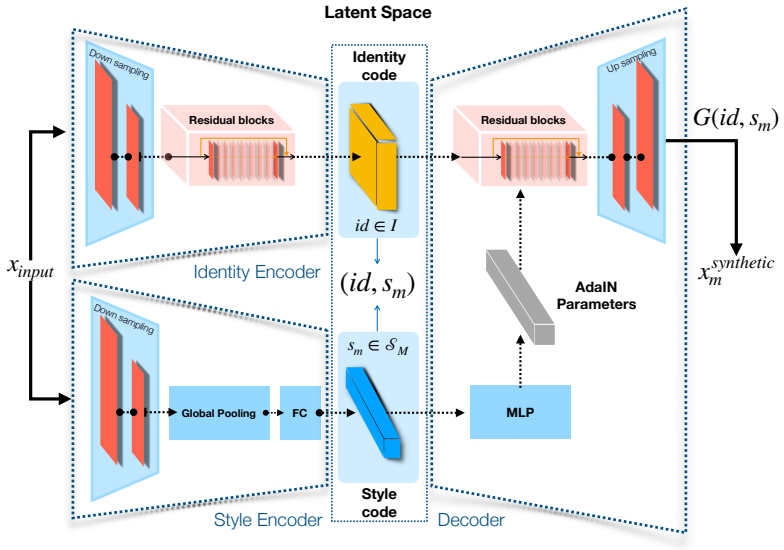


FIGURE 5.2: The auto-encoder architecture incorporates three networks. An *identity encoder*, which extracts a domain-shared face identity latent code id from x_{input} and consists of convolutional layers followed by several residual blocks. A *style encoder*, which extracts a domain-specific spectral information latent code s_m from x_{input} and consists of convolutional layers followed by a global average pooling layer and a last fully connected layer. A *decoder*, which reconstructs the image from prior identity and style code (id, s_m) generating $x_m^{synthetic}$.

Hence, we proceed to approximate the joint distribution via the latent space of the following two phases, *within-domain reconstruction* phase and *cross-domain translation* phase, respectively.

5.3.3 Network Architecture

Within-domain reconstruction phase

First, the identity latent code and style latent code are extracted from the input images x_{vis} and x_{thm}:

$$E_{\mathcal{V}}(x_{vis}) = (id_{vis}, s_{vis}) \text{ and } E_{\mathcal{T}}(x_{thm}) = (id_{thm}, s_{thm}). \tag{5.2}$$

The identity code and the style code come from the identity encoder and the style encoder, respectively drawn in Figure 5.2. The identity encoder consists of downsampling convolutional layers followed by several residual blocks to obtain the identity code, whereas the style encoder comprises of downsampling convolutional layers followed by global average pooling layer and a fully connected (FC) layer to output the style code.

Then, given the embedding of Equation (5.2), the face is reconstructed via the decoder (generator),

$$G_{\mathcal{V}}(id_{vis}, s_{vis}) = x_{vis}^{rec} \text{ and } G_{\mathcal{T}}(id_{thm}, s_{thm}) = x_{thm}^{rec}, \tag{5.3}$$

in order to learn the latent space for the specific face. Here, $\mathcal{M} \in \{\mathcal{V}, \mathcal{T}\}$ represents the domain, $E_{\mathcal{M}}$ denotes the factorized identity code and style code auto-encoder, $G_{\mathcal{M}}$ is the underlying decoder, and x_{vis}^{rec} and x_{thm}^{rec} are the corresponding reconstructed images. The decoder generate image from input identity and style code. Specifically, it processes the identity code by a set of residual blocks and produces the output image by several upsampling and convolutional layers. Meanwhile, residual blocks embedded into the decoder are equipped of Adaptive Instance Normalization (AdaIN) [61] layers whose parameters are dynamically generated by multilayer perceptron (MLP) from the style code.

The objective of LG-GAN is to learn the global *image reconstruction* mapping for a fixed $m \in \{vis, thm\}$, *i.e.,*

$$x_m \rightarrow x_m^{rec}, \tag{5.4}$$

while preserving facial identity features and allowing for a non-identity shift through *latent reconstruction* between

$$id_m \rightarrow id_m^{rec} \quad \text{and} \quad s_m \rightarrow s_m^{rec} \tag{5.5}$$

and forcing

$$s_{m-noise} \rightarrow s_m, \tag{5.6}$$

where, (id_m^{rec}, s_m^{rec}) are part of the extraction

$$(\ E_{\bar{\mathcal{M}}}(G_{\bar{\mathcal{M}}}(id_m, s_{\bar{m}-noise})), E_{\mathcal{M}}(G_{\mathcal{M}}(id_{\bar{m}}, s_{m-noise}))\ ),$$

respectively, and $s_{m-noise}$ is randomly drawn from a prior normal distribution in order to learn the associated style distribution. $\bar{\mathcal{M}}$ and $\bar{m}$ represent opposite domains.

Cross-domain translation phase

In the domain translation phase, image-to-image translation is performed by swapping the encoder-modality (*i.e.,* spectrum) with the alternate modality of the input image and imposing an explicit supervision on the style domain transfer functions $E_{\mathcal{V}}(x_{thm}) = (id_{thm}, s_{vis-noise})$ and $E_{\mathcal{T}}(x_{vis}) = (id_{vis}, s_{thm-noise})$, and then using $G_{\mathcal{V}}(id_{thm}, s_{vis-noise})$ and $G_{\mathcal{T}}(id_{vis}, s_{thm-noise})$ to produce the final output image $x_{vis/thm}^{fake}$ in the target spectrum. This is formalized as follows.

$$\begin{array}{rccc} \Theta_{t\rightarrow v} : & \mathcal{T} & \rightarrow & \mathcal{V} \\ & x_{thm} & \mapsto & x_{vis}^{fake} = G_{\mathcal{V}}(E_{\mathcal{V}}(x_{thm})); \end{array} \tag{5.7}$$

$$\begin{array}{rccc} \Theta_{v\rightarrow t} : & \mathcal{V} & \rightarrow & \mathcal{T} \\ & x_{vis} & \mapsto & x_{thm}^{fake} = G_{\mathcal{T}}(E_{\mathcal{T}}(x_{vis})). \end{array} \tag{5.8}$$

Consequently, $\Theta_{t\rightarrow v}$ and $\Theta_{v\rightarrow t}$ are the functions that synthesize the corresponding visible $(t \rightarrow v)$ and thermal $(v \rightarrow t)$ faces. Finally, LG-GAN learns the spectral conditional distribution $p_{\mathcal{V}|\mathcal{T}}(x_{vis}^{fake}|x_{thm})$ and $p_{\mathcal{T}|\mathcal{V}}(x_{thm}^{fake}|x_{vis})$ through a guided latent generation, where both these conditional distributions overcome

the fact that we do not have access to the joint distribution $p_{\mathcal{V},\mathcal{T}}(x_{vis}, x_{thm})$. Indeed, the method is able to generate, as an alternative, the joint distributions $p_{\mathcal{V},\mathcal{T}}(x_{vis}^{rec}, x_{thm}^{fake})$ and $p_{\mathcal{V},\mathcal{T}}(x_{vis}^{fake}, x_{thm}^{rec})$, respectively. We aim to learn the translation using neural networks, and this thesis will focus on Equation (5.7), where thermal face images are translated into realistic synthetic visible face images.

5.3.4 LG-GAN Training

LG-GAN is trained with the help of objective functions that include adversarial and bi-directional reconstruction losses as well as conditional, perceptual, identity, and semantic losses. We investigate the impact of each loss *w.r.t.* both biometrics and visual results and then propose an efficient combination. Furthermore, we use the VGG-19 [111] architecture which, when trained on a specific dataset, can be used to extract relevant features prior to applying the loss functions.

Adversarial Loss. Images generated during the translated phase through Equations (5.7) and (5.8) must be realistic and not distinguishable from real images in the target domain. Therefore, the objective of the generators, Θ, is to maximize the probability of the discriminator **Dis** making incorrect decisions. The objective of the discriminator **Dis**, on the other hand, is to maximize the probability of making a correct decision, *i.e.,* to effectively distinguish between real and fake (synthesized) images.

$$\begin{aligned}\mathcal{L}_{GAN}^{t\to v} = \mathbb{E}_{x_{vis}\sim p_{\mathcal{V}}}[log(\mathbf{Dis}_{\mathcal{V}}(x_{vis}))] + \\ \mathbb{E}_{x_{thm}\sim p_{\mathcal{T}}}[log(1-\mathbf{Dis}_{\mathcal{V}}(\Theta_{t\to v}(x_{thm})))],\end{aligned}$$

$$\begin{aligned}\mathcal{L}_{GAN}^{v\to t} = \mathbb{E}_{x_{thm}\sim p_{\mathcal{T}}}[log(\mathbf{Dis}_{\mathcal{T}}(x_{thm}))] + \\ \mathbb{E}_{x_{vis}\sim p_{\mathcal{V}}}[log(1-\mathbf{Dis}_{\mathcal{T}}(\Theta_{v\to t}(x_{vis})))].\end{aligned}$$

The adversarial loss is denoted as follows.

$$\mathcal{L}_{GAN} = \mathcal{L}_{GAN}^{t\to v} + \mathcal{L}_{GAN}^{v\to t}. \tag{5.9}$$

Bi-directional Reconstruction Loss. Loss functions in the Encoder-Decoder network encourage the domain reconstruction with regards to both image reconstruction and latent space (identity+style) reconstruction.

$$\mathcal{L}_{rec}^{image} = \mathop{\mathbb{E}}_{x_m^{rec}; x_m\sim p_{\mathcal{M}}}[\|x_{vis}^{rec} - x_{vis}\|_1 + \|x_{thm}^{rec} - x_{thm}\|_1], \tag{5.10}$$

$$\mathcal{L}_{rec}^{identity} = \mathop{\mathbb{E}}_{id_m^{rec}; id_m\sim p_{\mathcal{M}}}[\|id_{vis}^{rec} - id_{vis}\|_1 + \|id_{thm}^{rec} - id_{thm}\|_1], \tag{5.11}$$

$$\mathcal{L}_{rec}^{style} = \mathop{\mathbb{E}}_{s_m^{rec}; s_m \sim \mathcal{N}} [\|s_{vis}^{rec} - s_{vis}\|_1 + \|s_{thm}^{rec} - s_{thm}\|_1]. \tag{5.12}$$

Bi-directional refers to the reconstruction learning process between *image* → *latent* → *image* and *latent* → *image* → *latent* by the sub-network (a) and (b), respectively, depicted in Figure 5.1. Hence, the bi-directional reconstruction loss function is computed as follows:

$$\mathcal{L}_{rec} = \mathcal{L}_{rec}^{image} + \mathcal{L}_{rec}^{identity} + \mathcal{L}_{rec}^{style}. \tag{5.13}$$

Conditional Loss. Imposing a condition on the spectral distribution is essential and is the major difference from the baseline model [62]. Indeed, this allows for a translation that is conditioned the distribution of the target style code and, further, adds an explicit supervision on the final mapping $\Theta_{t \to v}$ and $\Theta_{v \to t}$. The conditional loss $\mathcal{L}_{cond}$ is defined as follows.

$$\begin{aligned} \mathcal{L}_{cond} = & \mathop{\mathbb{E}}_{s_{vis-noise}; s_{vis} \sim \mathcal{N}} \|s_{vis-noise} - s_{vis}\|_1 \\ & + \mathop{\mathbb{E}}_{s_{thm-noise}; s_{thm} \sim \mathcal{N}} \|s_{thm-noise} - s_{thm}\|_1. \end{aligned} \tag{5.14}$$

To improve the quality of the synthesized images and render them more realistic, we incorporate three additional objective functions, which are presented below.

Perceptual Loss. The perceptual loss $\mathcal{L}_P$ affects the perceptive rendering of the image by measuring the high-level semantic difference between synthesized and target face images. It reduces artefacts and enables the reproduction of realistic details [74]. $\mathcal{L}_P$ is defined as follows:

$$\begin{aligned} \mathcal{L}_P = & \mathop{\mathbb{E}}_{x_{vis}^{fake}; x_{vis} \sim p_{\mathcal{V}}} \|\phi_P(x_{vis}^{fake}) - \phi_P(x_{vis})\|_1 \\ & + \mathop{\mathbb{E}}_{x_{thm}^{fake}; x_{thm} \sim p_{\mathcal{T}}} \|\phi_P(x_{thm}^{fake}) - \phi_P(x_{thm})\|_1, \end{aligned} \tag{5.15}$$

where, ϕ_P represents features extracted by VGG-19, pre-trained on ImageNet. As a result, $\mathcal{L}_P$ is ensuring that the image represents properly a face.

Identity Loss. The identity loss $\mathcal{L}_I$ is responsible for preserving identity-specific features during the image reconstruction phase and, therefore, encourages the translated image to preserve the identity content of the input image. $\mathcal{L}_I$ is defined as follows:

$$\begin{aligned} \mathcal{L}_I = & \mathop{\mathbb{E}}_{x_{vis}^{rec}; x_{vis} \sim p_{\mathcal{V}}} \|\phi_I(x_{vis}^{rec}) - \phi_I(x_{vis})\|_1 \\ & + \mathop{\mathbb{E}}_{x_{thm}^{rec}; x_{thm} \sim p_{\mathcal{T}}} \|\phi_I(x_{thm}^{rec}) - \phi_I(x_{thm})\|_1, \end{aligned} \tag{5.16}$$

where, ϕ_I denotes the features extracted from the VGG-19 network pre-trained on the large-scale VGGFace2 dataset.

Semantic Loss. The semantic loss $\mathcal{L}_S$ guides the texture synthesis from thermal to visible domain and imparts attention to specific facial details.

A parsing network is used to detect semantic labels and to classify them into 19 different classes which correspond to the segmentation mask of facial attributes provided by CelebAMask-HQ [90]. We apply semantic face parsing to images in our datasets. A few examples are shown in Figure 5.3. $\mathcal{L}_S$ is defined as follows.

$$\begin{aligned}\mathcal{L}_S = &\mathop{\mathbb{E}}_{x_{vis}^{fake}; x_{vis}\sim p_{\mathcal{V}}} \|\phi_S(x_{vis}^{fake}) - \phi_S(x_{vis})\|_1 \\ &\mathop{\mathbb{E}}_{x_{thm}^{fake}; x_{thm}\sim p_{\mathcal{T}}} \|\phi_S(x_{thm}^{fake}) - \phi_S(x_{thm})\|_1,\end{aligned} \tag{5.17}$$

where, ϕ_S is the parsing network, providing corresponding parsing class label.

FIGURE 5.3: Example of face parsing results guided by the 19-class semantic label, when applied to images in the ARL-MMFD [52] and ARL-VTF [114] datasets.

Total loss. The overall loss function for the proposed LG-GAN is denoted as follows:

$$\min_{E_{\mathcal{V}},E_{\mathcal{T}},G_{\mathcal{V}},G_{\mathcal{T}}} \max_{\mathbf{Dis}} \mathcal{L}(E_{\mathcal{V}}, E_{\mathcal{T}}, G_{\mathcal{V}}, G_{\mathcal{T}}, \mathbf{Dis}) =$$

$$\lambda_{GAN}\mathcal{L}_{GAN} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{cond}\mathcal{L}_{cond} + \lambda_P\mathcal{L}_P + \lambda_I\mathcal{L}_I + \lambda_S\mathcal{L}_S. \tag{5.18}$$

Implementation Details. We implement the proposed LG-GAN framework in Pytorch by adapting MUNIT and designing the architecture for the modality-translation task. We note that we omit their proposed domain-invariant perceptual loss as well as the style-augmented cycle consistency. We train LG-GAN until convergence. The initial learning rate for Adam optimization is 0.0001 with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. For all experiments, the batch size is set to 1 and, based on empirical analysis, the loss weights are set to $\lambda_{GAN} = 1$, $\lambda_{rec} = 10$, $\lambda_{cond} = 35$, $\lambda_P = 15$, $\lambda_I = 20$ and $\lambda_S = 10$.

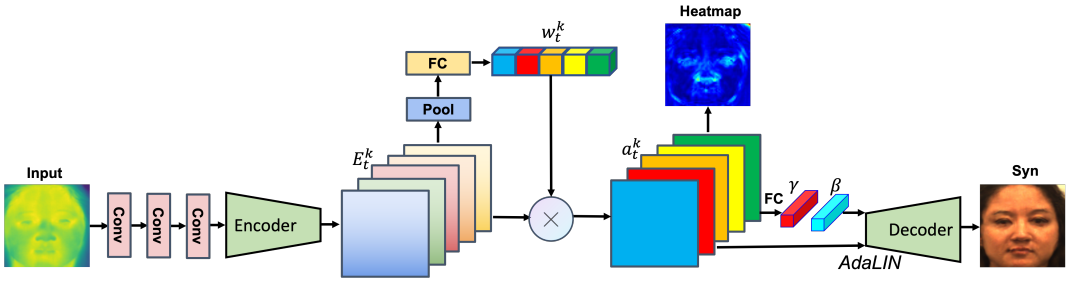


FIGURE 5.4: The architecture of the proposed AG-GAN and AG-GAN+ methods for thermal to visible image translation. The attention map in AG-GAN is generated by multiplying the learned attention weights with the feature maps obtained after encoder comprising downsampling bottlenecks. Such attention weights are obtained by inputting the GAP and GMP logits to an auxiliary classifier modulated by the CAM loss. The attention map in AG-GAN+ is generated by applying the squeeze-excitation (SE) module with no explicitly designed loss function to learn the attention weights. The decoder is formed by a series of upsampling bottlenecks coupled with AdaLIN parameters.

5.4 AG-GAN and AG-GAN+

In an attempt to enhance interpretability and investigate the specific facial features learned (across spectra) by a GAN during training (particularly by the generator and discriminator) for spectrum translation, we propose AG-GAN and AG-GAN+, two generic attention-guided generative adversarial networks designed for thermal-to-visible spectral translation. While AG-GAN is designed to encode thermal face images into attention maps learnt with *supervised* attention weights from an auxiliary-domain classifier, the AG-GAN+ version uses Squeeze and Excitation to generate in an unsupervised manner attention weights serving to build the attention map.

5.4.1 Baseline Model

Our method is inspired by the U-GAT-IT work [79] and the overall architecture is illustrated in Figure 5.4. Our approach involves notable differences, such as a complete redesign of the architecture to operate in a supervised learning framework with paired thermal-visible facial images. Unlike U-GAT-IT, we propose a novel method for learning attention feature maps from unsupervised attention weights. In addition, we introduce a set of new loss functions, including the identity loss, which allows us to better constrain the mapping space and improve performance. Overall AG-GAN encompasses three networks dedicated to the generation task, namely (i) encoder, (ii) attention module and (iii) decoder.

The main contributions of this work include the following.

- We design explainable generative adversarial networks based on attention feature map learning.
- We showcase the attention maps for both, generator and discriminator, and provide related quantitative and qualitative results.

- We compare AG-GAN and AG-GAN+, *w.r.t.* supervised and unsupervised attention learning. We offer extensive ablation studies and visualization results to validate the effectiveness of the proposed approaches from both, the controlled and the uncontrolled studies.

5.4.2 Formalization

Let $\mathcal{T}$ and $\mathcal{V}$ be the thermal and visible domains. Given an input thermal face image $x_{thm} \in \mathbb{R}^{H \times W \times C_{in}}$ and an output visible face image $x_{vis} \in \mathbb{R}^{H \times W \times C_{out}}$, the thermal-to-visible translation model can be described as:

$$\begin{array}{rccc} \Theta_{t \to v} & : & \mathcal{T} & \to & \mathcal{V} \\ & & x_{thm} & \mapsto & x_{vis}^{fake} = G_v(E_t(x_{thm})), \end{array} \tag{5.19}$$

where $\Theta_{t \to v}$ consists of an encoder E_t, a decoder G_v and an auxiliary-domain classifier $\eta_{\{\mathcal{T} \text{or} \mathcal{V}\}}$ which serves as generating attention weight for AG-GAN. Note that, AG-GAN+ does not include this classifier, as the attention weights are produced by a Squeeze and Excitation network (placed between the encoder and decoder).

Consequently, Equation 5.19 is the function producing the final output image x_{vis}^{fake} in the visible domain. Here, H, W, C_{in} and C_{out} are the height, width, input channel number and output channel number, respectively.

Let $x \in \{(x_{thm}, x_{vis}), (x_{thm}, x_{vis}^{fake})\}$ denote a sample pair conditioned on an input thermal image x_{thm}. Furthermore, the discriminator D_v is adopted to determine whether x is genuine or fake. In particular, (x_{thm}, x_{vis}) and $(x_{thm}, x_{vis}^{fake})$ denote the genuine and fake pairs, respectively.

5.4.3 Network Architecture

Generator:

Encoder. Given an input thermal image x_{thm}, we first use a 7×7 convolutional layer H_0 to transform an input image space into a high-dimensional feature space:

$$F_0 = H_0(x_{thm}). \tag{5.20}$$

Here, H_0 refers to a composite function of three different operations including convolution, instance normalization and ReLU. This operation was preceded by performing a reflection padding to keep the dimensional size unchanged. Then, we apply a sequence of down-sampling operations:

$$F_i = H_i(F_{i-1}), \tag{5.21}$$

where F_i represents the intermediate feature maps after the i-th down-sampling operation, for all $i \in \{1, ..., K\}$ with $K \in \mathbb{N}^*$.

Here, H_i is the same composite function as H_0, but with the purpose of halving the dimension and doubling the channel number. To further enhance

the feature embedding, we apply a series of residual blocks H_{R_j}:

$$F_j = H_{R_j}(F_{j-1}), \tag{5.22}$$

where F_j, for all $j \in \{K+1, ..., M\}$ with $M > K$, denotes the intermediate feature maps after performing feature enhancement including the j-th residual block, which has two 3×3 convolutional layers with the same output channel numbers and a skip-connection.

Attention module.

AG-GAN attention map. Given the embedding of Equations (5.21) and (5.22), we define the *encoder feature map* $E_t^k(x_{thm})$ as the k-th activation map from the encoder output F_M. In particular, we note $E_t^{k_{ij}}(x_{thm})$ as the value of activation map at (i, j). An auxiliary-domain classifier is later introduced to learn the weight w_t^k of the k-th feature map for the thermal domain. Thus, training is driven by both, *global average pooling* and *global max pooling*, viz. σ providing:

$$\eta_{\mathcal{T}}(x_{thm}) = \sigma\left(\sum_k w_t^k * \sum_{i,j} E_t^{k_{ij}}(x_{thm})\right). \tag{5.23}$$

In other words, $\eta_{\mathcal{T}}(x_{thm})$ expresses the probability that x_{thm} comes from the thermal domain. Finally, benefits from w_{thm}^k provide salient thermal domain specific *attention feature maps* that can be illustrated as follows:

$$a_t(x_{thm}) = w_t * E_t(x_{thm}). \tag{5.24}$$

Thereby giving rise to the proposed domain translation function

$$\Theta_{t \to v}(x_{thm}) = G_v(a_t(x_{thm})), \tag{5.25}$$

where we aim to learn the translation using neural networks.

AG-GAN+ attention map. Regarding AG-GAN+, the attention feature maps a_t are generated via the *squeeze-excitation* module that consists of squeeze and excitation operations. Mathematically, the squeeze operation can be described by

$$S_t = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} E_t(x_{thm})(i, j). \tag{5.26}$$

Here, H and W refer to the height and width of the encoded feature map $E_t(x_{thm})$ and (i, j) is the corresponding element. The channel-wise representation S_t is generated by applying the global average pooling per channel. For the excitation operation, S_t is fed into two sequential fully connected (FC) layers joined by the ReLU, which can described as follows,

$$w_t = \sigma\left(FC_1(ReLU(FC_2(S_t)))\right). \tag{5.27}$$

Here, w_t refers to the attention weights computed by SE. Next, we compute the resulted attention feature maps, see Equation 5.24.

Note that the difference between AG-GAN and AG-GAN+ lies in the generation of attention weights w_t. The AG-GAN *supervises* the weights by the auxiliary classifier associated with the CAM loss, whereas the AG-GAN+ generates the weights by squeeze-excitation operations without employing auxiliary loss function, hence, the inception of *unsupervised* learning. The heatmap is generated by averaging the channel dimension of the attention feature map a_t.

Decoder. The decoder's G_v purpose is to transform a high-dimensional feature space into an output image space. It comprises several residual blocks followed by up-sampling blocks. Here, residual blocks are instrumental in embedding features, while up-sampling convolution blocks generate target x_{thm}^{fake} visible domain images from the associated embedded features. Similarly to LG-GAN decoder, we further constrain the residual blocks with an extended Adaptive Layer-Instance Normalization (AdaLIN). AdaLIN combines both advantages of Adaptive instance normalization (AdaIN) and Layer Normalization (LN) by helping the AG-GAN(+) model to bring more flexibility in facial features generation control, *w.r.t.* shape and textures.

Discriminator:

The discriminator D_v performs a binary classification by determining whether the given pairs (x_{thm}, x_{vis}) and (x_{thm}, x_{vis}^{thm}) are genuine or fake. This is further enhanced by constructing two different scales of PatchGAN [68] discriminators that output resulting feature maps of 6×6 and 30×30, respectively. We adopt the same attention maps used by the generator and embed them into the discriminator for both AG-GAN and AG-GAN+.

5.4.4 AG-GAN and AG-GAN+ Training

Reconstruction Loss. We utilize the pixel-wise $\mathcal{L}_1$ loss function to measure the similarity between target visible face image x_{vis} and synthesized visible face image $\Theta_{t \to v}(I_t) = \tilde{I}_v$ at the pixel level:

$$\mathcal{L}_1 = \|x_{vis} - x_{vis}^{fake}\|_1. \tag{5.28}$$

The objective of $\mathcal{L}_1$ loss function is to minimize the difference at the low-level features, which may not lead to the preservation of high-level features such as identity.

While the reconstruction loss $\mathcal{L}_1$ is the same used in LG-GAN training, the identity loss differ.

Identity Loss. We utilize a pre-trained ArcFace recognition network [29] trained on MS-Celeb-1M dataset, noted ϕ_R, to extract facial feature embedding, measure the cosine similarity and compute the identity loss function as:

$$\mathcal{L}_{ID} = 1 - < \phi_R(x_{vis}), \phi_R(x_{vis}^{fake}) > . \tag{5.29}$$

Here, $\phi_R(x_{vis})$ and $\phi_R(x_{vis}^{fake})$ denote the normalized feature embedding. The identity loss differs from the previous identity loss used in LG-GAN. Indeed, an ablation study demonstrated that coupled with perceptual loss, the biometric performances decreased. Therefore, we opted to use ArcFace as identity loss, mainly due to its state-of-the-art performance in various FR benchmarks, such as the Labeled Faces in the Wild (LFW) dataset, which gained it popularity at the time of this thesis.

Perceptual Loss. As introduced in LG-GAN training, we utilize the perceptual loss to enhance the image quality:

$$\mathcal{L}_P = \|\phi_P(x_{vis}) - \phi_P(x_{vis}^{fake})\|_1. \tag{5.30}$$

Here, ϕ_P denotes the perceptual network constructed from the VGG-19. The overall loss function is the combination of aforementioned loss functions, along with the adversarial loss function.

CAM Loss. AG-GAN includes attention maps which are produced by attention weights learnt with an auxiliary-domain classifier $\eta_{\mathcal{T}/\mathcal{V}}$. This encourages AG-GAN at focusing on most relevant regions of a facial image while, translating the spectrum. To guide the classifier at distinguishing two domains, namely thermal $\mathcal{T}$ and V visible, we use a class activation mapping (CAM) loss function [154]. The CAM loss is conventionally used for image classification tasks and aims to identify the most important regions of an image that contribute to a particular class prediction.

The CAM loss results from the auxiliary classifier and can be described as:

$$\mathcal{L}_{CAM}^{\Theta_{t\to v}} = -(\mathbb{E}_{x\sim\mathcal{T}}[log(\eta_{\mathcal{T}}(x))] + \mathbb{E}_{x\sim\mathcal{V}}[1 - log(\eta_{\mathcal{T}}(x))]) \tag{5.31}$$

$$\mathcal{L}_{CAM}^{D_v} = \mathbb{E}_{x\sim\mathcal{V}}[\eta_{\mathcal{V}}(x)^2] + \mathbb{E}_{x\sim\mathcal{T}}[(1 - \eta_{\mathcal{V}}(G_v(x)))^2]. \tag{5.32}$$

Note that for AG-GAN+, no auxiliary classifier was used to learn the attention weights, thus no CAM loss was required.

Implementation Details. We use LSGAN[1] for training the AG-GAN by setting weight of $\mathcal{L}_1$, $\mathcal{L}_{ID}$, $\mathcal{L}_P$ and $\mathcal{L}_{CAM}^{\Theta_{t\to v}}$ to be 100, 1, 30, and 1000, respectively. The default number of epochs used in our training was 200. A batch size of 1 with the Adam optimizer was used. During the inference process, we

[1]LSGAN adopts the least squares loss function for the discriminator.

experimented with models generated at different epochs and selected epoch 90 as our final model.

5.5 Experiments

5.5.1 Datasets

To verify the effectiveness of the proposed methods, experiments were conducted on the controlled ARL-MMFD [52] and ARL-VTF [114] datasets, as well as the uncontrolled SpeakingFaces [2] datasets.

ARL-MultiModal Face dataset [52] (ARL-MMFD) contains visible, thermal (*i.e.,* LWIR), and polarimetric face images of over 60 subjects and includes variations in both expression and distance of acquisition. We only use visible and thermal images for our experiment at one particular stand-off distance: 2.5m. The first 30 subjects are used for testing and evaluation, and the remaining 30 subjects are used for training. The images in this dataset are already aligned and cropped.

ARL-VTF dataset [114] represents the largest academic collection of paired visible and thermal face images acquired in a time-synchronized manner. We follow the established evaluation protocol, which assigns 295 subjects for training and 100 subjects for testing and evaluation. We select the baseline gallery and probe subjects without glasses, named G VB0- and P TB0-, respectively. Furthermore, we align and process the images based on the provided eyes, nose and mouth ground truth landmarks.

SpeakingFaces dataset [2] (SF) consists of 142 subjects, where 100 subjects were used for training and the remaining subjects were used for testing. Following the established protocol, 5,400 and 2,268 thermal-visible image pairs were utilized for training and testing, respectively. Each subject was captured under 9 different poses, making it suitable for evaluating thermal-to-visible image translation under pose variations.

We here note that at the time, when the LG-GAN work was conducted, the ARL-VTF dataset was just released for public research purposes, alleviating the scarcity of the ARL-MMFD dataset, however the SF dataset was not accessible. The community switched from the ARL-MMFD to the ARL-VTF version, which is why in the work AG-GAN(+) we present experimentation with ARL-VTF.

5.5.2 Evaluation and Comparison

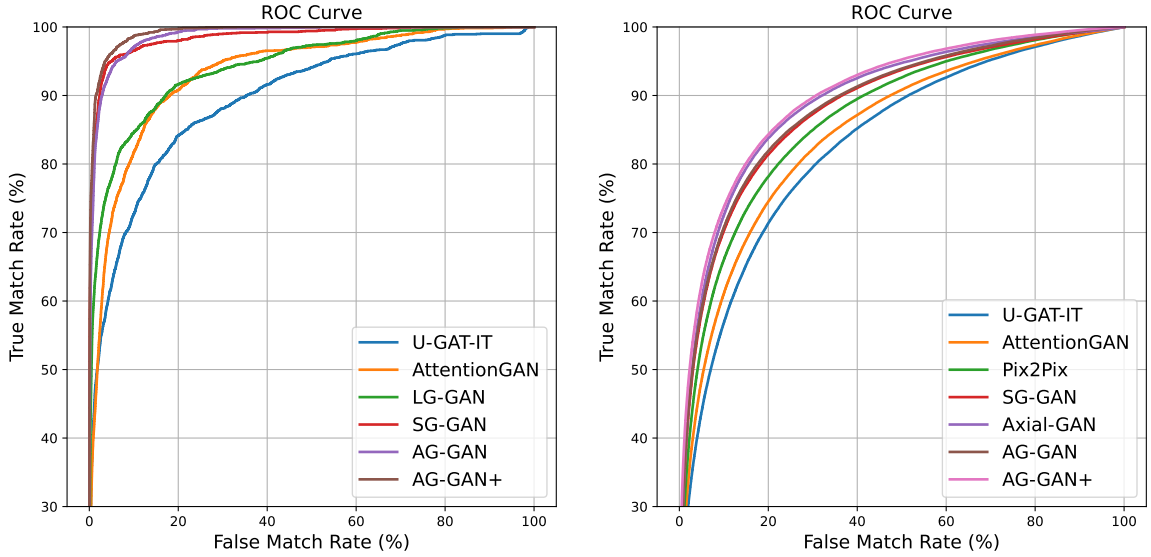
Tables 5.1, 5.2 and 5.3 show quantitative comparison between the proposed LG-GAN, AG-GAN and AG-GAN+ methods against other state-of-the-art methods, on three datasets respectively. Results, as introduced in Section 2.3.4, are expressed *w.r.t.* AUC and EER scores which measure the face matching accuracy, as well as PSNR and SSIM which assess the image quality. Note that we do not have access to the baseline code of other referenced methods from

Table 5.1, whereas Tables 5.2 and 5.3 reported results obtained by our own experimentation.

TABLE 5.1: Comparison of LG-GAN with other synthesis-based approaches on the ARL-MMFD dataset.

Method	AUC (%)	EER (%)
GAN-VFS [147]	79.30	27.34
AP-GAN [33]	84.16	23.90
AP-GAN (GT) [33]	86.08	23.13
Multi-stream GAN [148]	85.74	23.18
SAGAN [31]	91.49	15.45
SG-GAN [21]	93.08	14.24
Multi-AP-GAN [34]	90.74	18.20
Multi-AP-GAN (GT) [34]	92.72	16.05
LG-GAN (ours)	**93.99**	**13.02**

LG-GAN outperforms all other methods on the ARL-MMFD dataset (see Table 5.1) and is competitive with other methods on the ARL-VTF dataset (see Table 5.2). However, AG-GAN and AG-GAN+ outperform LG-GAN with a significant gap of performance. This could be a consequence of using another *identity loss* function which has more impact on discriminating biometrics features during the synthesis process (an ablation study on the impact of different identity loss function is further explored in Section 5.5.3). As a result, AG-GAN and AG-GAN+ achieve significantly higher true match rates across different false match rates showcased in Figure 5.5a.



(A) ARL-VTF dataset. (B) SF dataset.

FIGURE 5.5: ROC results of proposed algorithms and existing works

TABLE 5.2: Comparison of AG-GAN(+) and LG-GAN with other synthesis-based approach on ARL-VTF dataset.

Method	AUC↑	EER↓	PSNR↑	SSIM↑
U-GAT-IT [79]	89.58	17.84	14.49	0.65
AttentionGAN [130]	93.35	13.41	14.34	0.64
LG-GAN (ours)	96.96	5.94	16.81	0.68
SG-GAN [21]	98.52	5.07	17.27	0.71
Axial-GAN [63]	99.05	4.98	-	-
AG-GAN (ours)	98.74	5.56	17.39	0.70
AG-GAN+ (ours)	**99.26**	**4.30**	**17.58**	**0.72**

Notably, AG-GAN and AG-GAN+ allow for an increased similarity in generating accurately visible-like images from thermal inputs, while preserving identity well and incorporating finer details (see Figure 5.6). In contrast, U-GAT-IT, AttentionGAN and LG-GAN appear to include more artifacts.

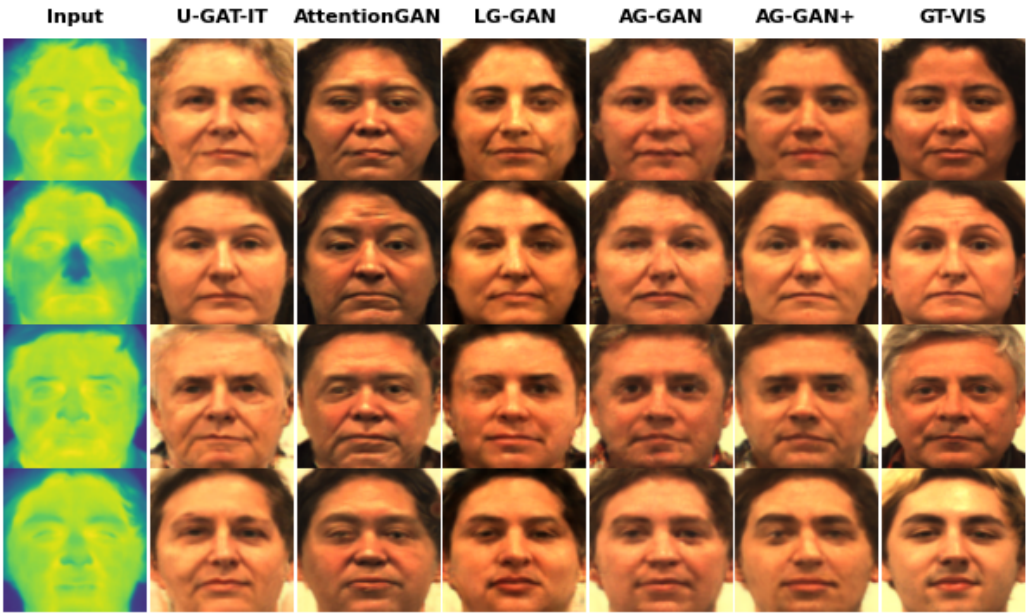


FIGURE 5.6: Comparison of qualitative results of proposed algorithms with existing works on ARL-VTF dataset.

Further, for evaluating thermal-to-visible translation which deals with facial pose variation, we experiment with the the SF dataset. Table 5.3 demonstrates the effectiveness of AG-GAN and AG-GAN+ methods over other state-of-the art methods. This is corroborated by the use of face verification and image quality metrics, and ROC curves are depicted for selected algorithms in Figure 5.5b. In addition, a qualitative comparison between the methods is shown in Figure 5.7. Both proposed methods are able to synthesize high-fidelity visible face images under pose variations, while preserving the identity well.

TABLE 5.3: Comparison of AG-GAN and AG-GAN+ with other synthesis-based approach on SF dataset.

Method	AUC↑	EER↓	PSNR↑	SSIM↑
U-GAT-IT [79]	83.0	24.48	19.05	0.71
CUT [2]	83.62	23.59	20.51	0.67
AttentionGAN [130]	84.69	22.92	19.21	0.71
Pix2Pix [68]	86.82	20.99	20.29	0.72
CycleGAN [2]	86.97	20.70	20.34	0.67
SG-GAN [21]	88.41	19.23	20.33	0.72
Axial-GAN [63]	89.51	17.90	**21.15**	0.69
AG-GAN (ours)	89.86	17.68	20.82	0.74
AG-GAN+ (ours)	**90.53**	**17.20**	21.01	**0.75**

FIGURE 5.7: Comparison of qualitative results of proposed algorithms with existing works on SF dataset.

5.5.3 Ablation Study

Loss functions are crucial in providing realistic face images as well as preserving biometrics features. To illustrate the impact of the loss functions included in our experimentation, on both (i) visual quality and (ii) FR scores, we conduct an ablation study using the ARL-VTF dataset. This dataset is known as the largest academic collection of paired visible-thermal faces and thus provides more accurate results (see Section 3.3).

Impact of loss functions

Based on the work of SG-GAN [21], LG-GAN expanded the use of *perceptual loss*, *identity loss* and *semantic loss*, which play the role of enforcing both (i-ii) challenges.

Table 5.4 highlights the need to use pertinent loss functions since the first intuitive observation has to do with the low performance of direct matching between thermal and visible face images that can be explained in terms of the

TABLE 5.4: Face verification performance, image quality, and impact of different loss functions on ARL-VTF [117] dataset. $\mathcal{L}_{base}$ represents the method including the adversarial (5.9) bi-direction reconstruction (5.13) and conditional (5.14) losses, while $\mathcal{L}_P$, $\mathcal{L}_I$, $\mathcal{L}_{P+I}$ and $\mathcal{L}_{P+I+S}$ are the perceptual (5.15), identity (5.16) and semantic (5.17) losses added to the original $\mathcal{L}_{base}$ training.

	ARL-VTF Dataset		
	AUC (%)	EER (%)	SSIM
Direct comparison	54.80	46.31	0.3739
$\mathcal{L}_{base}$	92.21	15.88	0.6049
$\mathcal{L}_P$	92.79	14.24	0.6129
$\mathcal{L}_I$	92.98	13.01	0.6101
$\mathcal{L}_{P+I}$	92.15	15.36	0.6136
$\mathcal{L}_{P+I+S}$ = LG-GAN	**96.96**	**5.94**	**0.6787**

lower AUC of 54.80 and SSIM of 0.3739. This illustrates once more the modality gap.

In an attempt to overcome this modality gap, the first baseline experiment $\mathcal{L}_{base}$, without visual quality optimization, boosts the AUC and SSIM scores to 92.21 and 0.6049, respectively. However, we note that from Figure 5.8, generated results are rather blurry and not realistic. By adding additional loss functions to $\mathcal{L}_{base}$, $\mathcal{L}_P$ (Equation (5.15)) and $\mathcal{L}_I$ (Equation (5.16)), namely the perceptual and identity losses, the SSIM score only marginally increases. Jointly, $\mathcal{L}_{P+I}$ improves the SSIM score further. Nevertheless, AUC and EER depict another trend. Individually, $\mathcal{L}_P$ and $\mathcal{L}_I$ increase the score while $\mathcal{L}_{P+I}$ slightly decrease to AUC to 92.15. We point out that *perceptual* contents of faces are about pixel level, whereas *biometrics* information are related to features level. Although the perceptual loss increases the reconstructed facial similarity (see $\mathcal{L}_P$ in Figure 5.8), the associated AUC in Table 5.4 reports lower performances in terms of AUC and EER. And conversely, the identity loss provides higher biometric results than the perceptual loss, but the similarity (SSIM) is lower. Finally, when adding the semantic loss $\mathcal{L}_S$ (Equation (5.17)) as a synthesis guidance, LG-GAN is able to generate more realistic images with less artefact along with a higher similarity to the visible ground truth and accurate for face recognition. However, semantic parser masks are not stable during the training, which does not allow the model to achieve higher results.

Choice of the identity loss

Towards designing an efficient CFR system, *identity* is the first concern. In contrast, image quality is considered being secondary to biometric performance (in which we place emphasis on). This motivates the change of the identity loss version, from LG-GAN to AG-GAN(+), with an alternative network able to maintain competitive biometric results when associated to the perceptual loss (realistic face). Therefore, in AG-GAN(+) we introduced $\mathcal{L}_{ID}$ (Equation (5.29)), an identity loss relying on ArcFace [29] recognition network. Further details about ArcFace can be found in Section 2.3.3.

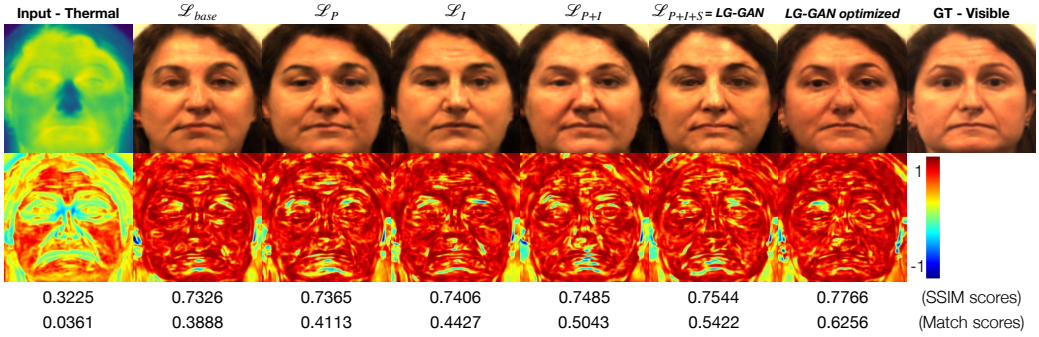


FIGURE 5.8: A visualization strategy to understand the impact of different loss functions on the visual quality as well as the matching scores. The top row shows samples generated by individual and combined loss functions, while the bottom row illustrates the SSIM scores as well as the SSIM *similarity and difference* of two images in different scenarios: *GT-Visible* against *Input-Thermal*/$\mathcal{L}_{base}$/$\mathcal{L}_P$/$\mathcal{L}_I$/$\mathcal{L}_{P+I}$/*LG-GAN*/*LG-GAN optimized*.

TABLE 5.5: Ablation study on the impact of identity (ID) loss with ARL-VTF dataset using AG-GAN.

Method	AUC↑	EER↓	PSNR↑	SSIM↑
w/o ID	97.66	7.78	17.02	0.69
w/ MobileFaceNet	97.59	7.57	16.94	0.69
w/ ArcFace = $\mathcal{L}_{ID}$	**98.74**	**5.56**	**17.39**	**0.70**

The choice of using such $\mathcal{L}_{ID}$ identity loss relies on the following observation. We investigate the use of different FR networks to extract feature embedding and compute identity loss. As seen in Table 5.5, using ArcFace [29] to derive the feature embedding for identity loss computation results in higher performance than MobileFaceNet [23], as well as the framework without adopting the identity loss. Admittedly, applying a FR loss does not always result in improved matching performance. Therefore, it is critical to choose a suitable identity loss to ensure maximum identity similarity between target visible and synthesized visible images.

Attention map as a guidance

While LG-GAN considers semantic content as a synthesis guidance, AG-GAN uses attention maps. AG-GAN guides the translation process to highlight salient regions of the face and disregard less significant ones by using attention map generated by the help of auxiliary classifiers, which allow to differentiate between the thermal and visible spectra. The generator and discriminator are equipped with attention maps that emphasize semantically significant regions, simplifying the process of translating facial images from one spectrum to another. The generator's attention map directs attention towards areas that are crucial in distinguishing between the thermal and visible domains, whereas the discriminator's attention map aids in refining the model by focusing on the contrast between real and fake images in the target-visible spectrum.

By exploiting the information from the auxiliary classifier, the CAM loss assists the generator and discriminator to place interest on similarity across spectrum. In following LG-GAN assumption, which state that *identity* features is shared by images of a subject acquired in different spectra, the CAM loss is responsible to preserve identity discriminative features during translation.

The Table 5.6 conduces an ablation study to understand the effectiveness of the CAM loss embedded on the generator and discriminator, namely GCAM loss $\mathcal{L}_{CAM}^{\Theta_{t\to v}}$ and DCAM loss $\mathcal{L}_{CAM}^{D_v}$. The use of CAM loss can increase the face matching accuracy. Here, "w/o DCAM" and "w/o GCAM" refers to the settings, where no CAM loss is applied to the discriminator and generator, respectively. "w/o GDCAM" refers to the setting where no CAM loss is applied to both, generator and discriminator. However, we do not observe a correlation between face matching accuracy and perceived image quality when analyzing the CAM loss, which confirm their action on only identity preservation aspect (thus validate LG-GAN assumption).

TABLE 5.6: Ablation study on the impact of CAM loss with ARL-VTF dataset.

Method	AUC↑	EER↓	PSNR↑	SSIM↑
AG-GAN	**98.74**	**5.56**	17.39	0.70
w/o DCAM	97.19	7.09	17.00	0.69
w/o GCAM	97.93	6.30	17.67	0.70
w/o GDCAM	97.74	7.29	**17.81**	**0.71**

5.6 Discussion

5.6.1 Latent code visualization (LG-GAN)

Understanding the latent code is critical for LG-GAN, as it aims to elicit identity-specific information while ignoring spectrum induced information. A disentangled latent space is produced using an identity encoder and a style encoder that decomposes the input image into an identity code and a style code; see Figure 5.2. As discussed earlier, the style code represents the spectral information and drives the domain translation, without adversely affecting the identity information. However, here we seek to explore whether identity is explicitly encoded in the latent space. Towards this goal, we visualize the identity code, id_m, directly after the encoding step $E_{\mathcal{M}}(x_m)$; then, by up-scaling the code to the target image size, we determine the pertinent pixels that are responsible for the identity information in the latent space. This is visualized in Figure 5.9. We observe that facial features around eyes, nose, mouth and hair have been encoded. Moreover, identity codes – id_{vis} and id_{thm} – extracted from both spectra also highlight the same visual information. This is consistent with the partially shared latent space assumption adopted in LG-GAN and made by Huang *et al.* [62].

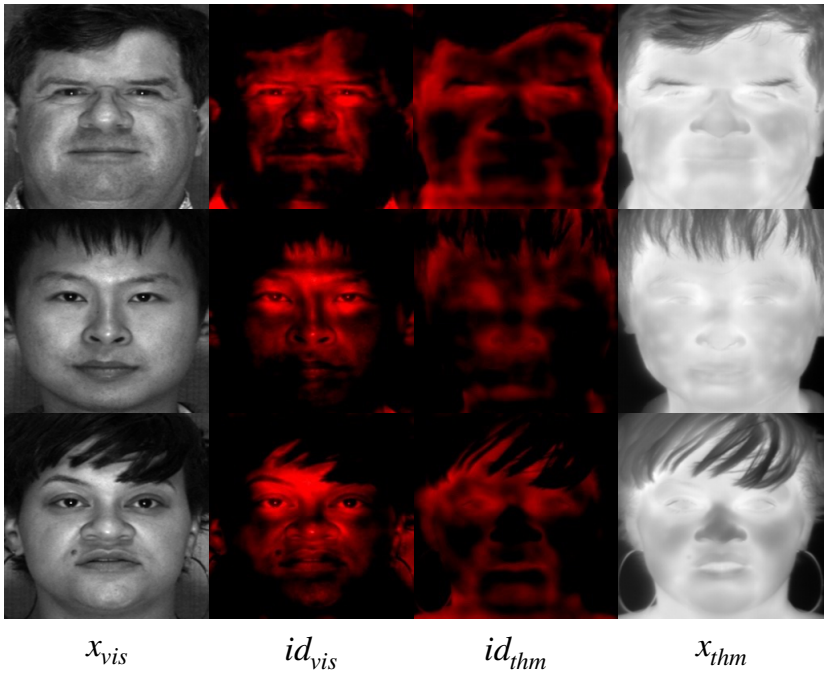


FIGURE 5.9: Visualization of identity codes, id_{vis} and id_{thm}, extracted from $E_{\mathcal{V}}(x_{vis})$ and $E_{\mathcal{T}}(x_{thm})$, respectively.

5.6.2 Attention map visualization (AG-GAN)

Towards interpreting the generation process from thermal to visible images, we visualize the attention maps embedded in the generator. As seen in Figure 5.10, features are generally activated around the salient facial regions including nose, mouth and eyes, consistent with the identity code showcased in LG-GAN. Other notable regions such as hair are also likely to be activated. However, it is worth pointing out that the skin regions are not activated in the resulting attention maps. Overall, as indicated the generator network tends to focus on salient facial features that are discriminative across subjects, namely the biometric "identity" information. The skin region, on the other hand, appears to share more similarity in texture, thereby not being accentuated by the generator. In contrast, the discriminator-*attention maps* focus on distinguishing the skin region difference between the synthesized and ground-truth visible face images. This could be explained by the fact that features from salient facial regions are well synthesized, hence considering to be genuine by the discriminator. Thus, the generator and discriminator are unlikely to compete against each other on these regions to distinguish between genuine and fake images.

We show that the attention maps entail highly similar representations during the entire generation process. Figure 5.11 reveals the attention maps produced at different test epochs for a single subject. It becomes evident that attention maps at different stages demonstrate highly consistent representations in salient facial regions including eyes, nose and mouth. This clearly validates that the visual attention is consistent across the entire training process. Note that with the increase of epochs, more visually pleasing images can be obtained.

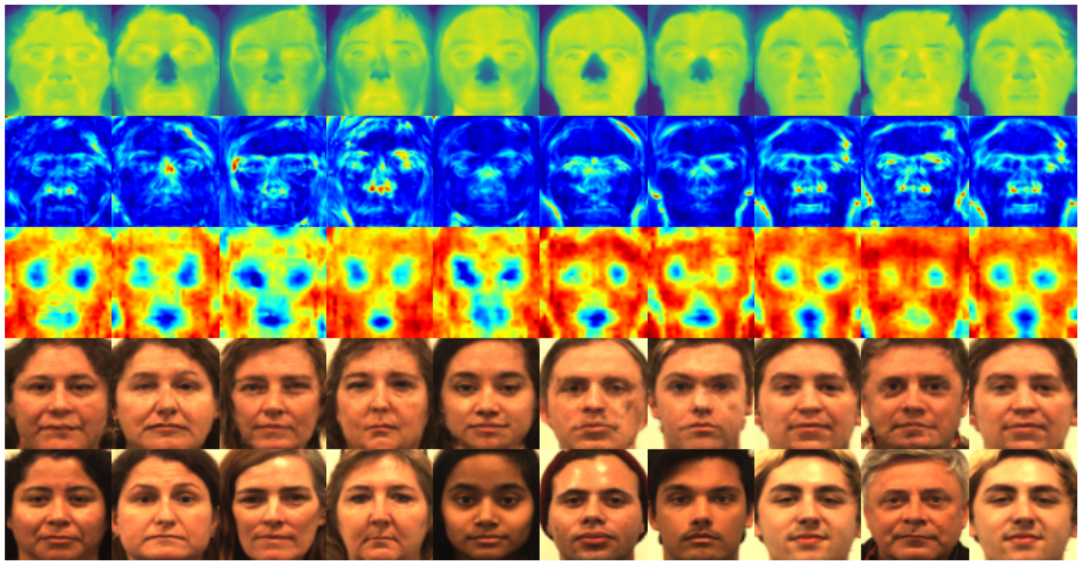


FIGURE 5.10: **Transparency and Interpretability.** Examples of attention maps produced by the generator and discriminator on ARL-VTF dataset using AG-GAN. The images from top to bottom rows are: thermal, generator attention map, discriminator attention map, synthesized visible and ground-truth visible face images.

5.6.3 Supervised vs. Unsupervised Attention

The attention weights from the proposed AG-GAN method were learned by the CAM loss through auxiliary classifiers, while the attention weights from AG-GAN+ were learned by the *squeeze-excitation* operation. Here, the *supervised* and *unsupervised* learning were defined based on whether the attention weights are directly learned by a loss function.

Figure 5.12 shows the heatmaps generated by the squeeze-excitation. Compared to Figure 5.10, where the attention maps are more localized, the AG-GAN+ generates the attention maps that are more globally distributed. This is challenging for the model explainability, as there are no consistent patterns observed. However, the attention maps computed by the generator still clearly reveal the structure information also consistent with the identity code provided by LG-GAN, as similar areas are highlighted.

5.7 Summary

In this Chapter, we proposed a latent-guided generative adversarial network (LG-GAN) that explicitly decomposes an input image into an identity code and a style code. The identity code is learned to encode spectral-invariant identity features between thermal and visible image domains in a supervised setting. In addition, the identity code offers useful insights in explaining salient facial structures that are essential to the synthesis of high-fidelity visible spectrum face images. Experiments on two datasets suggest that our proposed LG-GAN achieves competitive thermal-to-visible cross-spectral face recognition accuracy,

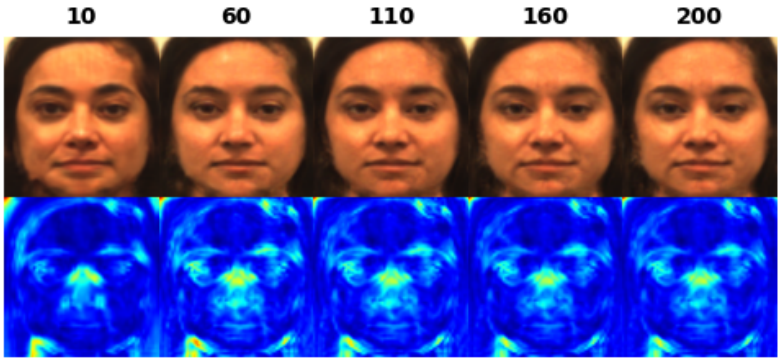


FIGURE 5.11: **Robustness and Consistency.** Examples of attention maps produced by the generator at individual test epochs on ARL-VTF dataset using AG-GAN. The images in the top and bottom rows are synthesized visible images and their corresponding attention maps.

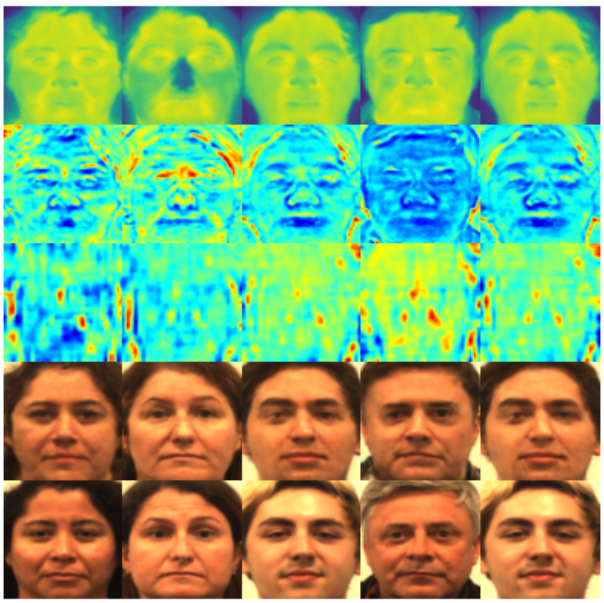


FIGURE 5.12: Examples of attention maps produced by unsupervised attention learning using AG-GAN+, ordered by thermal, generator and discriminator attentions, synthesized and ground-truth visible from ARL-VTF dataset.

while enabling explanations on salient features used for thermal-to-visible image translation.

Towards deepening the understanding of thermal-to-visible image translation, identity code representation has been enhanced with attention modules, giving rise to an attention-guided generative adversarial network (AG-GAN). Attention maps are extracted from the encoder in both, supervised and unsupervised manner and are fed into the AdaLIN-based decoder. By visualizing the learned attention maps, we showed that AG-GAN is capable of interpreting thermal-to-visible image translation. Additionally, AG-GAN+ takes advantage of squeeze and excitation operation for learning attention weight in unsupervised manner and demonstrates better cross-spectral face recognition results,

outperforming other methods. Finally, a commonality between LG-GAN and AG-GAN has to do with the use of specific loss functions that are critical to the faithful identity preservation, as well as the ability to generate realistic faces, thereby enabling the development of more accurate and robust CFR systems.

Chapter 6

Towards an end-to-end cross-spectral face recognition system

The task of thermal FR becomes even more arduous when scenarios of application involve recognizing individuals in unconstrained settings. In such context, *facial pose* and *image quality* can vary significantly. Additionally to such compounding factors, CFR systems often necessitate to analyze images captured at various distances, and hence image face resolutions, as well as images of individuals, who are randomly situated away from the camera (see Figure 6.1). Such *uncontrolled* acquisition of multi-scale (low) resolution thermal images inherently incorporates scarce detailed facial information, which might be insufficient for FR. Prior approaches have addressed CFR by considering a fixed resolution, necessitating that a subject stands at a precise distance from a given sensor during acquisition, which constitutes an impracticable scenario in real-life. Towards loosening this constraint, as well as recognizing faces from any (low) resolution of thermal images, we propose a unique model handling the dual computer vision tasks of *domain translation* and *super-resolution*.

In this Chapter, we address a practical real-world CFR setting, where we aim at comparing thermal face images of any (low) resolution to a gallery of high resolution visible face images by ANYRES. ANYRES generates high resolution visible images from low resolution thermal images, placing emphasis on *maintaining the cross-spectral identity*. Building upon the experimental findings and observations from the previous Chapter 5, we will design an *encoder-decoder* based architecture and leverage relevant loss functions to optimize the learning process. Such loss functions have demonstrated a consistent identity preservation across spectra. Section 6.1 highlights the critical need to design a model that is adaptable to different standoff distances in CFR scenarios, as encountered in real-world operations, while Section 6.2 revisits recent approaches and techniques addressed to perform *super-resolution*, as well as *domain translation*. Section 6.3 introduces the framework of the proposed ANYRES and Section 6.4 reports experimental results with main focus on presenting TFLD combined to ANYRES as a successive task towards designing an end-to-end CFR system. Finally, the versatility of the method is presented in Section 6.5. Specifically, we discuss the choice of resolutions for training, as well as the capability to process thermal faces with variation in pose.

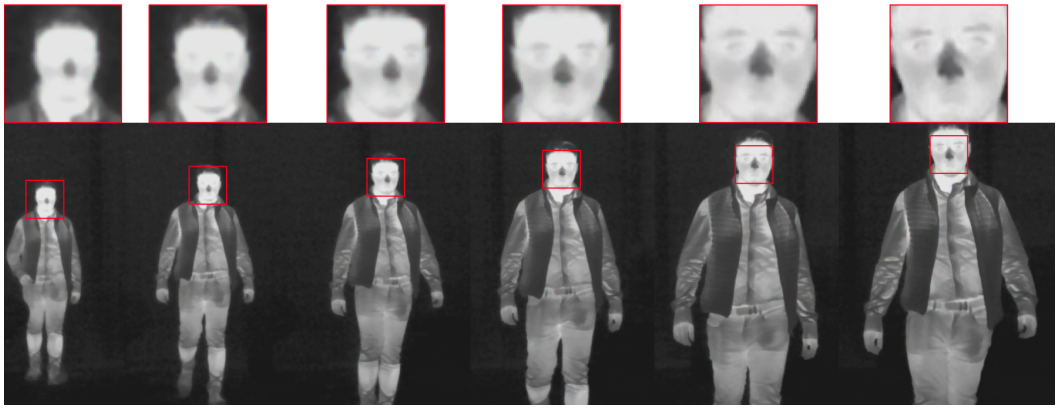


FIGURE 6.1: During operational applications, humans are randomly situated away from the camera and can therefore depict multi-scale (low) resolution thermal face images (resolution depends of the acquisition distance).

6.1 Introduction

Thermal sensors play a crucial role in detecting and recognizing humans in surveillance settings, specifically in the context of long-range distance acquisition or under adverse lighting conditions (low-light or night-time environments). However, thermal imaging does not provide detailed rendering of faces, hindering related FR systems. Therefore, generating visible-spectrum face images of High-Resolution (HR) based on associated thermal-spectrum face images of Low spatial Resolution (LR) is of particular pertinence in designing operational CFR systems [5], where for example a visible face image is compared to a face image acquired beyond the visible spectrum. Such generation process is referred to as Super Resolution (SR) or Hallucination, aiming to produce HR images based on single or sequential LR images. All existing solutions allow for SR from a *fixed input resolution* [63], rendering them completely impractical in real-life scenarios.

Motivated by the above, we here address the task of comparing thermal face images of any (low) resolution against a gallery of HR visible face images by designing a unique model handling dual computer vision tasks, *i.e., super resolution* and *domain translation*, streamlined to be more adaptive to ensure faithful cross-spectral identity preservation. In particular, we propose ANYRES, a novel model that allows for simultaneous face SR, as well as thermal-to-visible spectrum translation. We have placed emphasis on ANYRES being robust to any LR thermal inputs, while preserving the identity of depicted individual. Benefits from the simultaneous process are instrumental for avoiding accumulated errors and artifacts. ANYRES is equipped to bridge simultaneously the modality gap, as well as the resolution gap. In particular, a blurry, thermal LR face image is transformed into a sharp, realistic, HR visible-like face image. The designed network presents the advantage of preserving consistent biometric features across both, the LR/HR space, as well as the thermal/visible spectrum, allowing for comparison of super resolved images and a gallery of visible images,

using off-the-shelf FR algorithms. Furthermore, the proposed algorithm is suitable to real world operational scenarios in which humans are randomly situated away from the camera, resulting in multi-scale LR thermal face images. Deviating from the state-of-the-art, where resolution is generally fixed *w.r.t.* input, ANYRES is to the best of our knowledge the first framework to operate at any input resolution ranging from LR to HR.

The main contributions of this work include the following.

- We propose a novel supervised learning framework for CFR that performs simultaneously both, *domain translation* and *super-resolution*. Specific loss functions have been introduced, in order to enhance both, image quality as well as biometric feature preservation.
- We empower the network by learning to process a range of resolutions as inputs, while previous methods enabled only fixed-resolution inputs. Our mechanism is based on a resolution-*inter-dependency*, (i) taking advantage of pyramidal architecture, as features are perceived with multi-scale analysis and (ii) gating spectral encoded features with decoded super resolved features.
- We achieve state-of-the-art performance on four benchmark multi-spectral face datasets, with respect to *visual quality*, as well as *face recognition* scores.

6.2 Background

Large resolution discrepancy between visible and thermal sensors induces paired visible-thermal face datasets with images having significant resolution-gap [5]. Although existing CFR methods [20, 7] are based on GANs to simulate artificial visible-like facial images from thermal face images, they remain challenged by altering image resolution. Such methods predominantly focused on conditional-GANs, where multi-spectral paired facial samples were used for supervised learning. Consequently, Pix2Pix designed with UNet-like encoder-decoder architecture was adopted for its ability to learn conditional mapping from one domain to another. Further optimization was introduced to constrain *perceptual*-rendering [112], *identity*-preservation [151], and *semantic*-attribute guidance [21], between the synthesized visible face images and the target visible face images. Early works on reliable CFR systems in unconstrained environments include the work of Immidisetti *et al.* [63] that proposed the first study exploring factor resolution in the context of long-range surveillance systems. The associated AxialGAN is streamlined to accommodate CFR, in cases of images being captured at different distances from depicted humans. AxialGAN addresses simultaneously spectrum translation from thermal-to-visible, as well as face hallucination, however restricted to a fixed resolution of the input image. AxialGAN incorporated an axial-attention layer to capture long-range dependencies, included in both generator and discriminator networks.

6.3 ANYRES

We propose ANYRES, a GAN designed to address simultaneously both tasks, *domain translation*, as well as *super resolution* from varying-resolution inputs, while preserving identity. In particular, ANYRES tackles the problem of matching thermal face images of any LR against HR visible face images by (i) learning an end-to-end mapping between the thermal spectrum and the visible spectrum, and (ii) learning to handle input of any resolution.

6.3.1 Problem Formulation

We here consider the HR space, with cardinality $m \times n$, incorporating a visible domain $\mathcal{V}$ with visible face images $x_{vis} \in \mathbb{R}^{m \times n}$, and a thermal domain $\mathcal{T}$ with thermal face images $x_{thm} \in \mathbb{R}^{m \times n}$.

Domain translation phase In the domain translation phase, image-to-image translation is performed by learning an end-to-end non-linear mapping, denoted as $\Theta_{t \to v}$, between the thermal spectrum and the visible spectrum. This is formalized as follows:

$$\begin{array}{ccccc} \Theta_{t \to v} & : & \mathcal{T} & \to & \mathcal{V} \\ & & x_{thm} & \mapsto & x_{vis}^{synthetic}. \end{array} \tag{6.1}$$

$\Theta_{t \to v}$ represents the function that synthesizes the corresponding thermal face images into realistic synthetic visible face images $x_{vis}^{synthetic}$ in the HR space.

Super Resolution phase. Given the embedding of Equation (6.1), the network encapsulates the SR scalability as a simultaneous task. Therefore, we aim to learn a conditional generation function, where a thermal LR facial image $x_{thm}^{LR} \in \mathbb{R}^{\frac{m}{r} \times \frac{n}{r}}$ is also enhanced to the HR scale, providing a synthetic visible image $x_{vis}^{SR} \in \mathbb{R}^{m \times n}$ up-scaled by a $\times r > 0$ scale factor, via

$$x_{vis}^{SR} = \Theta_{t \to v}(x_{thm}^{LR}). \tag{6.2}$$

As elaborated above, thermal-to-visible FR based on GAN-synthesis, with the objective of being robust to any LR thermal inputs, aims to learn a unified function that, when applied to *any*-LR thermal image x_{thm}^{LR}, yields a higher-resolution super resolved (SR) visible image $x_{vis}^{SR} \in \mathbb{R}^{m \times n}$ with rich semantic and identity information. In this context, the contribution of ANYRES is the simultaneous learning of global interaction between both *domain translation* and *resolution* scalability through the enrichment of Equation (6.1) by Equation (6.2). To be specific, for all scale factors $0 < r \leq m$, the method $\Theta_{t \to v}$ is designed to learn neural networks by considering (x_{thm}, x_{vis})-paired facial images and minimizing specific loss functions (supervised setting).

6.3.2 Baseline Model

Towards learning how to process any resolution as input, without having to estimate said resolution, we designed ANYRES to be endowed with a U-shape pyramidal architecture. It relies naturally on a multi-scale analysis. The overall architecture is illustrated in Figure 6.2.

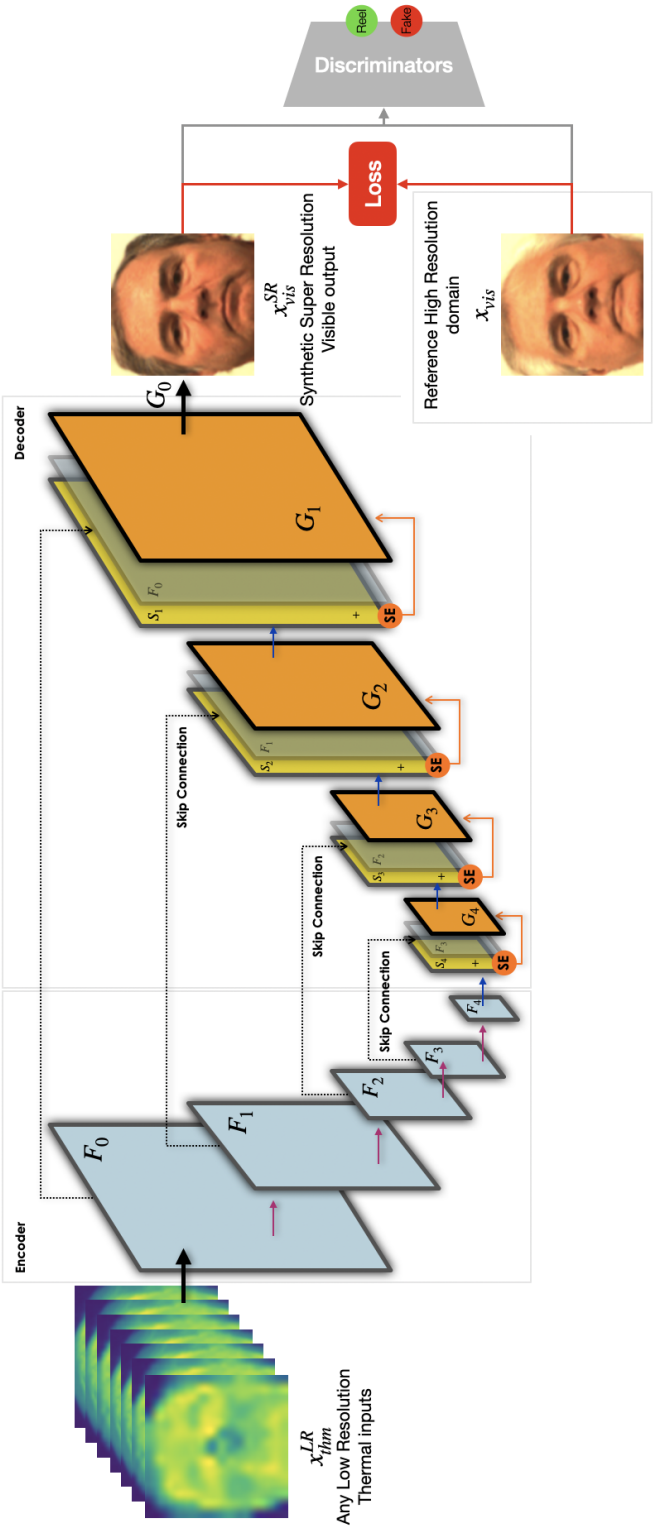


FIGURE 6.2: **Training of ANYRES.** The generator accepts any (low)-resolution thermal face x_{thm}^{LR} as input. It comprises an encoder-decoder bridged by skip connections and gated by Squeeze and Excitation (SE) blocks, which play the role of gate modulator and enable resolution-wise relationships towards bringing a flexible control for balancing encoded features with decoded super resolved features. The discriminators are aimed at distinguishing real images x_{vis} from generated synthetic ones x_{vis}^{SR}.

We model our function $\Theta_{t\to v}$ using a U-net architecture. The generator consists of an encoder-decoder structure with skip connections between domain specific encoder and decoder. Considering the larger discrepancy between the images resulted from LR and HR spaces, we introduce *Squeeze-and-Excitation* [51] (SE) blocks, which play the role of gate modulator after each skip connection. Such strategy enables channel-wise relationships and brings a flexible control for balancing encoded features with decoded super resolved features.

Generator:

Cross-Resolution Interaction. During training time, the network is fed simultaneously with batches of a wide range r-scale factors of (low) resolution thermal images. In the case of a fixed r, the model is able to super resolve images from $\frac{m}{r} \times \frac{n}{r}$ to $m \times n$ scale of space (*i.e.,* fixed LR input unlike any LR input). In what follows, we refer to a model trained with one scale factor as *mono-resolution*, whereas a model trained with several scale factors is denoted *multi-resolution*.

Encoder. The encoder extracts multi-resolution features in parallel and fuses them repeatedly during the learning stage, in order to generate high-quality SR-*representations* with rich semantic/identity information.

Given a LR thermal input image x_{thm}^{LR}, we first use a layer H_0 to transform a LR input image space into a high-dimensional feature space

$$F_0 = H_0(x_{thm}^{LR}). \tag{6.3}$$

Here, H_0 refers to a composite function of two successive *Convolution-BatchNormalization-ReLU* layers. Then, we apply a sequence of operations

$$F_i = H_i(\text{Pool}(F_{i-1})), \tag{6.4}$$

where F_i represent the intermediate encoded feature maps after the i-th operation, for all $i \in [1, K]$ with $K \in \mathbb{N}^*$. Here, H_i is the same composite function defined in Equation (6.3), and Pool denotes a max pooling operation where the most prominent features of the prior feature map are preserved.

Decoder. The decoder aims at transforming a high-dimensional feature space into a SR output image in the visible spectrum. Hence, the generative task towards the super resolved images is started from the deep level (U bottleneck),

$$G_K = H_K(SE(C(F_{K-1}, S_K(F_K)))). \tag{6.5}$$

Then sequentially incremented, for all $i \in [1, K-1]$, by

$$G_i = H_i(SE(C(F_{i-1}, S_i(G_{i+1})))), \tag{6.6}$$

ending by the generation of the SR image x_{vis}^{SR} through *Convolution-Tanh* layers

$$G_0 = x_{vis}^{SR}. \tag{6.7}$$

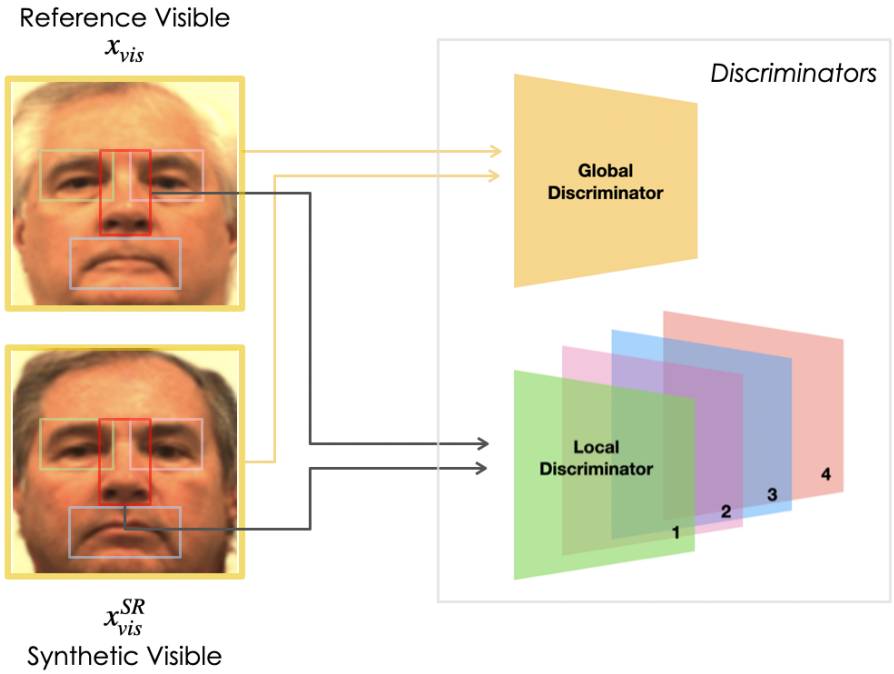


FIGURE 6.3: **Global and Local discriminators**. While the global discriminator, applied on the whole image, is instrumental for the generator to synthesize photo-realistic HR images, the local discriminators, denoted by L_1, L_2, L_3 and L_4, focus on areas located around eyes, nose and mouth, respectively. They are designed to focus on generated details of cross-spectral biometric features.

While S refers to upsampling operation by factor 2 followed by *Convolution-BatchNormalization-ReLU* layers, C concatenates all channels from the skip connection F_{i-1} with the up-sampled S_i layers. Finally, G_i represent the decoded intermediate feature maps after the i-th operation preceded by Squeeze and Excitation SE.

Discriminator:

In an adversarial learning, ANYRES is complemented by global and local discriminators, named $\mathbf{Dis}_{\text{global}}$ and $\mathbf{Dis}_{\text{local}}$ respectively, further depicted in Figure 6.3. The former helps the generator to synthesize photo-realistic HR image, whereas the latter is focused on subtle facial details and benefits from local inherent attention to capture faithful biometric features during the generation.

Global Discriminator. We adopt the multi-scale discriminator, which enables generation of realistic images with refined details. Hence, $\mathbf{Dis}_{\text{global}}$ is responsible of performing a binary-classification by distinguishing a super resolved image x_{vis}^{SR} from a real image x_{vis}.

Local Discriminator. To synthesize biometric-realistic semantic content, we focus on discriminant areas relevant for identification. Such regions are represented by the same cropping area (see Figure 6.3) between the images x_{vis} and

x_{vis}^{SR}, respectively named $x_{vis-ROI,i}$ or $x_{vis-ROI,i}^{SR}$, with $i \in [0, 4]$. Thus, the local discriminator $\mathbf{Dis}_{\text{local}}$ extends the design to the independent discriminators L_i, placing attention on every single facial fine detail and benefit from local inherent attention, in order to capture faithful biometric features during the generation.

6.3.3 Training

The learning process of ANYRES is driven by an efficient combination of objective functions inspired from LG-GAN and AG-GAN in Chapter 5, that pave the way to control the synthesis process at both *pixels* and *features* levels.

Adversarial loss. Images generated through Equation (6.1) must be realistic. Therefore, the objective of the generator is to maximize the probability of the discriminators making incorrect decisions. The objective of the discriminators, on the other hand, is to maximize the probability of making a correct decision, *i.e.,* to effectively distinguish between real and synthesized images. The global $\mathcal{L}_{GAN}^{Global}$ and local $\mathcal{L}_{GAN}^{Local}$ loss functions are part of the adversarial training and defined as follows:

$$\mathcal{L}_{GAN} = \mathcal{L}_{GAN}^{Global} + \mathcal{L}_{GAN}^{Local}. \tag{6.8}$$

Conditional loss. Imposing a condition on the spectral distribution is essential for generating images within the target spectrum. The conditional loss (known as L1 loss) is defined as follows.

$$\mathcal{L}_{cond} = \mathop{\mathbb{E}}_{x_{vis}; x_{vis}^{SR} \sim p_{\mathcal{V}}} \|x_{vis}^{SR} - x_{vis}\|_1. \tag{6.9}$$

Perceptual loss. The perceptual loss $\mathcal{L}_P$ affects the perceptive rendering of the image (ensuring they are representing faces) by measuring the high-level semantic difference between synthesized and target face images. It reduces artefacts and enables the reproduction of realistic details. $\mathcal{L}_P$ is defined as follows:

$$\mathcal{L}_P = \mathop{\mathbb{E}}_{x_{vis}; x_{vis}^{SR} \sim p_{\mathcal{V}}} \|\phi_P(x_{vis}^{SR}) - \phi_P(x_{vis})\|_1, \tag{6.10}$$

where, ϕ_P represents features extracted by VGG-19, pretrained on ImageNet.

Identity loss. The identity loss $\mathcal{L}_I$ preserves the identity of the facial input and relies on a pre-trained ArcFace [29] recognition network to extract facial features embedding. This choice stems from the discussion related to Section 5.5.3. Then, the cosine similarity measure provides the identity loss function

$$\mathcal{L}_I = \mathop{\mathbb{E}}_{x_{vis}; x_{vis}^{SR} \sim p_{\mathcal{V}}} [1 - < \phi_I(x_{vis}), \phi_I(x_{vis}^{SR}) >], \tag{6.11}$$

where, ϕ_I denotes the features extracted from Arcface.

Attribute loss. The attribute loss $\mathcal{L}_A$ prevents attribute shift during spectrum translation. While age brings apparent information, gender relies on identity. Therefore, apparent age loss $\mathcal{L}_A^{Age}$ and gender loss $\mathcal{L}_A^{Gender}$ are defined as follows.

$$\mathcal{L}_A^{Age} = \mathop{\mathbb{E}}_{x_{vis}; x_{vis}^{SR} \sim p_{\mathcal{V}}} \|\phi_{Age}(x_{vis}^{SR}) - \phi_{Age}(x_{vis})\|_1, \tag{6.12}$$

$$\mathcal{L}_A^{Gender} = \mathop{\mathbb{E}}_{x_{vis}; x_{vis}^{SR} \sim p_{\mathcal{V}}} \|\phi_{Gender}(x_{vis}^{SR}) - \phi_{Gender}(x_{vis})\|_1, \tag{6.13}$$

where, ϕ_{Age} and ϕ_{Gender} are pre-trained models based on DeepFace facial attribute framework analysis [121]. Then, the attribute loss is denoted as follows.

$$\mathcal{L}_A = \mathcal{L}_A^{Age} + \mathcal{L}_A^{Gender} \tag{6.14}$$

Finally, jointly all loss functions bring to the fore realism during spectral translation, while avoiding blurriness introduced by any low scale of resolution from thermal image inputs. ANYRES relies on the combination of the aforementioned loss functions.

6.3.4 Implementation Details

ANYRES is implemented in PyTorch and uses Adam optimizer with an initial learning rate of 0.0002, and $\beta_1 = 0.5$, $\beta_2 = 0.999$. For all experiments, the batch size and the default number of epochs used are set to 4 and 100, respectively.

Images are first aligned with the eyes, nose and mouth key points by adopting the protocol of Anghelone *et al.* [8] and scaled to the HR size of 128×128. For the training phase, LR images are down-sampled from the HR thermal images with four different scale factors giving batch of images of size 128×128, 64×64, 32×32 and 16×16, respectively. Moreover, the training dataset is augmented by random sharpness, center cropping and horizontal flips.

Although facial alignment and re-scale operations are used to simulate different resolution at a given image size, the SR task needs to be addressed. These operations involve space-transformation, namely homography. However, such transformations have no positive effect in terms of *information-related intrinsic pixel enhancement*. Details and information that are missing cannot be recovered by simply re-scaling or aligning the image. Then, aligning the face and scaling the image do not fundamentally change the nature of the task from a SR problem to solely an image-to-image translation problem. Even if the scale operation increases the spatial resolution to a target size, there is a need to restore the information with SR techniques. In particular, faces captured in the thermal spectrum depict other representations than faces captured in the visible spectrum and such techniques attempt to find information in different parts of the face, in order to restore poor thermal face images into high resolved visible-like face images.

TABLE 6.1: Characteristics of datasets used for the experiments

Dataset	ARL-VTF	VIS-TH	SF	Tufts
Reference	[114]	[98]	[2]	[109]
Number of training subjects	295	40	100	50
Number of testing subjects	100	10	42	63

6.4 Experiments

6.4.1 Datasets

Towards evaluating the performance of our network, we train ANYRES on four benchmark multi-spectral face datasets, individually. To the best of our knowledge, these are the most popular datasets in the field. We follow the respective train-test splits, as suggested by the associated authors for comparison purposes. We summarize the datasets in Table 6.1.

ARL-VTF dataset [114] is considered as the largest collection of paired thermal and visible face images and includes 500,000 images captured with variations in expression, pose, and eyewear. Following the established evaluation protocol, 295 subjects were used for training and 100 subjects were used for testing.

VIS-TH dataset [98] was collected from 50 subjects, with 40 subject for training and 10 for testing, resulting a total of 2100 images. Variations include illumination, facial expression, change in head pose, and occlusions.

SF dataset [2] included 142 subjects captured under 9 different extreme poses. The first 100 subjects are used for training, and the remaining for testing.

Tufts dataset [109] consists of 113 subjects, including 39 males and 74 females under different modalities. Following the established protocol used in [42], 50 subjects were randomly selected for training and the remaining subjects were used for testing.

6.4.2 Evaluation and Comparison

Figure 6.4 and Table 6.2 highlight qualitative and quantitative comparison results of different methods. Results are reported in terms of (i) FR biometrics standards, we present the Area Under the Curve (AUC) and Equal Error Rate (EER) metrics related to the ArcFace-based FR matcher[1]; as well as (ii) image quality evaluated by the structural similarity index measure (SSIM)[2].

To validate the effectiveness of ANYRES, we implemented from provided code: *face hallucination*, *super resolution*, *domain translation* and related *leading* dedicated methods for comparison purpose, respectively named HiFaceGAN [143], SRGAN [89], Pix2Pix [68] and AxialGAN [63].

While ANYRES is designed to handle *any* resolution (shortened by ANYRES-multi), we trained other methods that had been originally designed for specific (mono) resolution. Note that ANYRES-mono is further included as an ablation study.

[1]Higher AUC indicates better performance, whereas lower EER is better.
[2]Score of 1 is the extreme case of comparing identical images.

TABLE 6.2: Quantitative comparison on four multi-spectral face datasets. Experimental results validate accuracy *w.r.t.* facial recognition, namely by AUC % and EER % scores, as well as visual quality, as reported by SSIM %. Bold indicates the best performance.

Res.	Method	**ARL-VTF dataset**			**VIS-TH dataset**			**SF dataset**			**Tufts dataset**		
		AUC	EER	SSIM	AUC	EER	SSIM	AUC	EER	SSIM	AUC	EER	SSIM
16 × 16	HiFaceGAN	63.44	40.97	43.59	67.65	41.35	37.64	69.65	35.54	49.60	57.27	45.26	32.53
	SRGAN	82.68	25.69	50.21	69.91	36.86	49.15	71.76	34.09	56.57	50.86	49.31	14.76
	Pix2Pix	84.07	23.70	55.13	69.43	39.57	52.26	73.93	32.27	59.86	59.77	43.11	32.46
	AxialGAN	84.45	22.92	59.74	72.14	33.82	53.82	77.61	29.12	64.69	59.65	43.10	38.79
	ANYRES-mono	90.18	17.68	61.78	73.75	32.74	53.35	77.63	28.81	**65.43**	63.59	40.30	38.83
	ANYRES-multi	**91.24**	**16.90**	**63.05**	**75.94**	**29.66**	**53.89**	**77.64**	**28.04**	65.34	**63.62**	**40.23**	**39.01**
32 × 32	HiFaceGAN	82.55	26.32	46.41	79.53	36.95	39.25	74.03	32.22	50.31	62.17	41.33	32.94
	SRGAN	95.26	12.02	56.50	84.86	22.04	54.01	82.81	24.79	64.20	58.44	44.14	23.25
	Pix2Pix	95.78	10.95	58.36	82.39	32.54	53.08	80.97	26.48	62.44	65.04	39.05	32.81
	AxialGAN	95.19	11.85	64.43	84.95	22.50	57.36	84.24	23.34	67.93	63.38	40.39	39.91
	ANYRES-mono	**98.88**	**5.23**	**66.76**	**88.91**	**18.51**	**57.91**	**85.80**	**21.68**	**68.44**	76.15	30.50	**40.79**
	ANYRES-multi	98.61	6.17	64.23	87.01	20.82	55.35	84.77	23.08	68.33	**77.22**	**29.09**	40.01
64 × 64	HiFaceGAN	86.74	21.79	48.55	82.06	34.56	41.40	77.46	29.24	55.21	63.09	40.81	36.73
	SRGAN	97.85	5.30	59.72	88.07	20.86	56.97	86.02	21.42	63.02	66.60	38.04	33.62
	Pix2Pix	97.93	6.90	60.48	86.23	28.52	54.16	85.35	22.27	63.31	65.41	38.89	33.67
	AxialGAN	97.22	9.19	66.39	88.09	20.88	57.78	86.14	21.49	67.33	64.23	39.69	41.31
	ANYRES-mono	**99.86**	**1.75**	**68.06**	**93.62**	**14.23**	**58.24**	**90.81**	**16.84**	**69.24**	**83.07**	**24.60**	**41.76**
	ANYRES-multi	99.42	4.02	67.05	89.57	18.08	57.43	86.24	21.24	68.85	80.74	26.41	41.38
128 × 128	HiFaceGAN	91.10	17.10	50.58	83.49	33.67	49.18	79.10	27.78	52.05	65.02	39.45	36.88
	SRGAN	98.61	5.14	59.28	88.16	20.18	57.14	87.47	20.27	67.07	67.47	37.58	39.08
	Pix2Pix	98.21	7.07	60.57	87.37	20.10	54.34	86.61	21.10	63.76	65.50	38.78	33.88
	AxialGAN	97.91	9.65	66.48	88.59	20.96	57.98	87.02	20.52	67.61	66.27	38.23	40.00
	ANYRES-mono	**99.88**	**1.26**	**68.61**	**93.65**	**14.05**	**58.46**	**91.39**	**15.87**	**69.68**	**83.09**	**24.53**	**41.43**
	ANYRES-multi	99.44	3.82	67.02	89.58	18.04	57.98	89.13	18.14	68.93	80.91	26.29	40.30

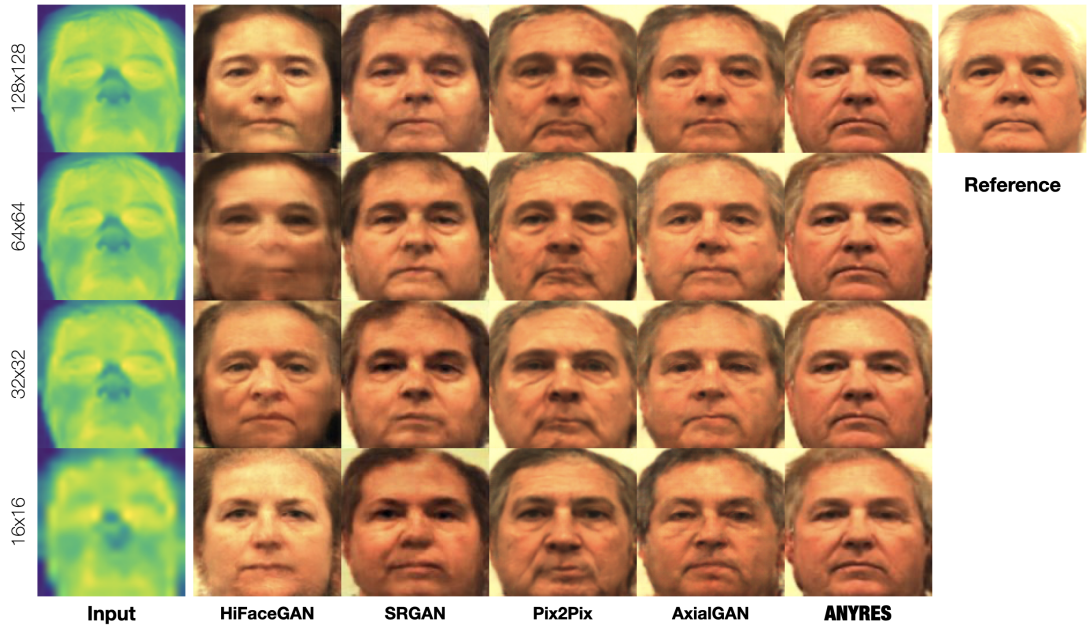


FIGURE 6.4: Qualitative results of HiFaceGAN, SRGAN, Pix2Pix, AxialGAN and the proposed ANYRES on the ARL-VTF dataset. We decrease the resolution in each row (re-scaled to 128×128). While previous methods are impaired to super resolve facial images for a given resolution by using one specific network for each resolution, our proposed ANYRES achieves a balance between realism and fidelity across resolutions with only one unified network.

All tested methods rely on adversarial training, nevertheless they differ in the way they ensure faithful cross-spectral identity. Our first observation is that ANYRES outperforms other methods across datasets for all resolutions *w.r.t.* FR performances, in *mono* or *multi* mode. More specifically, ANYRES-mono outperforms all other comparable methods, including ANYRES-multi, in higher resolution settings. As expected however, ANYRES-multi has a superior performance in lower resolutions (*i.e.,* 16x16). Moreover, from resolution 32×32 to 128×128, we notice that biometric performances are roughly the same across resolutions, which indicates the ability of identity-consistency through various resolutions. *W.r.t.* image quality, ANYRES depicts stable SR images across resolutions without artefacts. In almost all resolutions and datasets, it achieves either best or a nearly second best SSIM score as opposed to other methods. We note that CFR relies on biometric features rather than perceptual features (pixel scale), and therefore we consider image quality being secondary to biometric performance, which we place emphasis on.

Unexpectedly, SRGAN which is originally built to super-resolve an image within the same spectrum, has demonstrated competitive results that could surely be improved, in case that specific loss functions were added. Nevertheless, its design is not optimized for accumulating resolutions and this could be

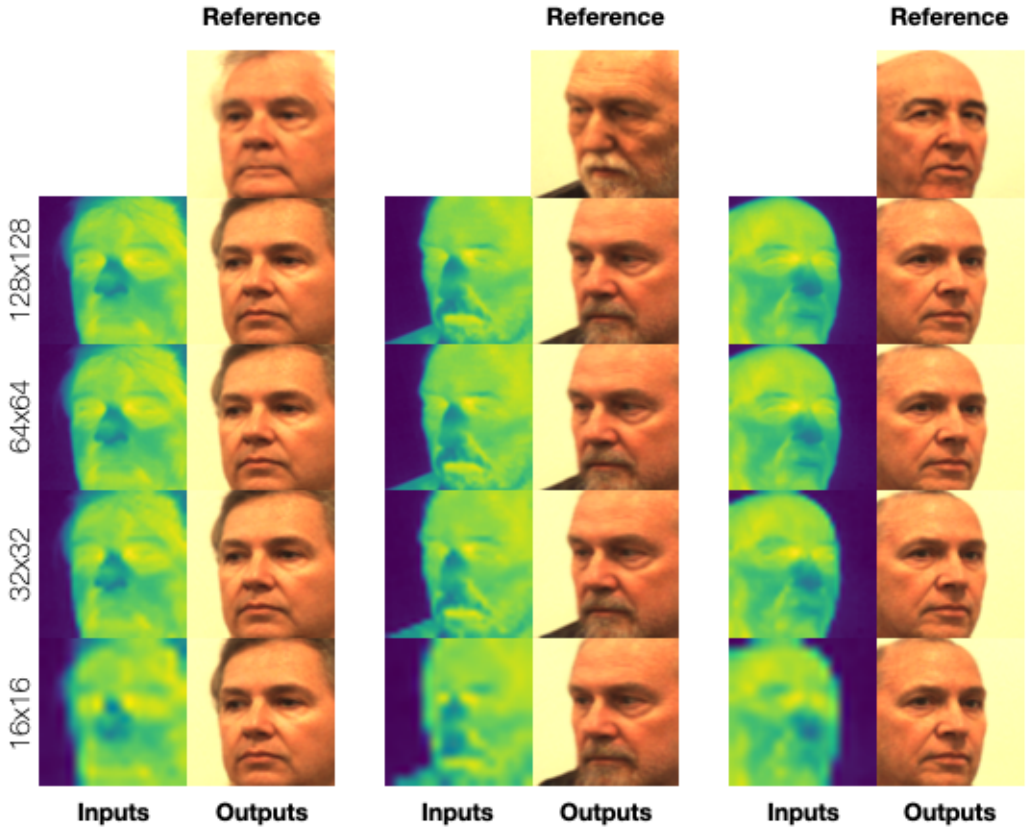


FIGURE 6.5: Qualitative results of ANYRES (pose-to-pose) on the *pose* subset of the ARL-VTF dataset. We decrease resolution in each row (re-scaled to 128×128).

explained by the residual blocks fashion processing. Finally, our approach significantly boosts the performances and demonstrates the ability to handle any resolution solely with one unified framework. Results presented on the ARL-VTF dataset significantly exceed the comparative scores. This gap is explained by the fact that, ARL-VTF is the largest thermal-visible paired face dataset publicly available, and includes over 500,000 images, unlike other datasets which contain around one hundred images.

With respect to observation in pose, we trained ANYRES on two multi-spectral facial benchmarks, namely SF and the subset "pose" of ARL-VTF. From a non frontal face as input, ANYRES is able to generate a synthetic face with the same facial pose. We name this process *pose-to-pose*. Note that both datasets contain (extreme) pose faces, which brings random variation during training. Further results confirm the ability of ANYRES to be operational in unconstrained-CFR systems. Figure 6.5 shows qualitative results whereas Table 6.3 reports quantitative results on four resolutions. The results presented for the "pose" subset of the ARL-VTF dataset are rather similar to those of the SF dataset, indicating that ANYRES performs consistently when confronted with profile face images. This suggests that the ANYRES model is capable of handling variations in pose, and therefore can be utilized in practical scenarios, where face images are captured from various angles.

TABLE 6.3: Quantitative comparison of ANYRES (pose-to-pose) on the subset *pose* of ARL-VTF dataset. Experimental results validate accuracy *w.r.t.* facial recognition, namely by AUC % and EER % scores, as well as visual quality, as reported by SSIM %.

ANYRES-multi	***pose*-ARL-VTF dataset** *pose-to-pose*		
Res.	AUC	EER	SSIM
16×16	84.28	22.99	65.96
32×32	90.75	15.05	68.12
64×64	91.99	13.45	68.22
128×128	91.96	13.51	68.19

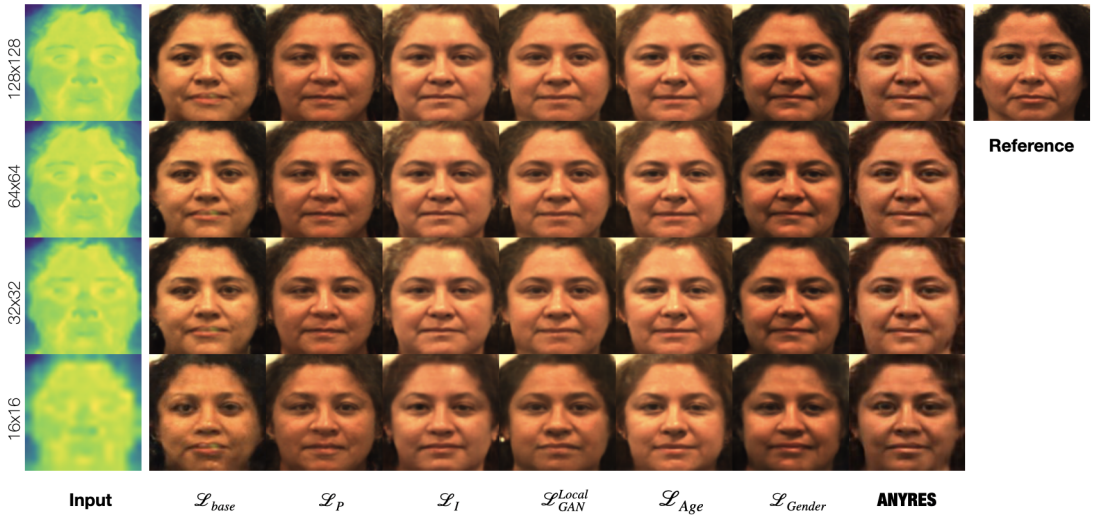


FIGURE 6.6: A visualization strategy, from the ARL-VTF dataset, to understand the impact of different loss functions (incrementally added from $\mathcal{L}_{base}$ to the whole combination giving ANYRES) on the visual quality ranging from high to low resolution. We decrease resolution in each row (re-scaled to 128×128).

We do not propose a direct comparison (matching) between thermal and visible face images across all protocols. The face matcher (ArcFace) is not trained to extract discriminative features from thermal images. An intuitive reasoning will induce that we would get mediocre results (this illustrates once more the modality gap that is depicted in Table 3.7).

Overall, our proposed method, ANYRES, achieves state-of-the-art performance on four benchmark multi-spectral face datasets, demonstrating its effectiveness in both "mono" and "multi" version.

6.4.3 Ablation Study

To illustrate the impact of *loss functions* included in ANYRES on face recognition scores, we conduct an ablation study using the ARL-VTF dataset. This dataset is known as the largest dataset on the field and Table 6.4 reports AUC and EER scores under different experimental configurations. Besides,

TABLE 6.4: **Ablation study.** Face Verification Performance (AUC and EER) and impact of different loss functions of ANYRES-multi on ARL-VTF dataset. $\mathcal{L}_{base}$ represents the method including global-adversarial and conditional losses, while $\mathcal{L}_P$, $\mathcal{L}_I$, $\mathcal{L}_{GAN}^{Local}$ and $\mathcal{L}_A$ are the Perceptual, Identity, Local and Attributes losses added to the original $\mathcal{L}_{base}$ training. The best score is indicated in bold.

input Res.	$\mathcal{L}$ added	**ARL-VTF dataset** AUC (%)	EER (%)
16×16	$\mathcal{L}_{base}$	86.70	20.49
	$+\mathcal{L}_P$	88.13	19.42
	$+\mathcal{L}_I$	90.62	16.90
	$+\mathcal{L}_{GAN}^{Local}$	91.03	16.94
	$+\mathcal{L}_A$ = ANYRES	**91.24**	**16.90**
32×32	$\mathcal{L}_{base}$	96.80	9.93
	$+\mathcal{L}_P$	97.26	9.07
	$+\mathcal{L}_I$	97.94	7.82
	$+\mathcal{L}_{GAN}^{Local}$	98.09	7.01
	$+\mathcal{L}_A$ = ANYRES	**98.61**	**6.17**
64×64	$\mathcal{L}_{base}$	97.53	8.45
	$+\mathcal{L}_P$	98.30	7.23
	$+\mathcal{L}_I$	98.83	5.63
	$+\mathcal{L}_{GAN}^{Local}$	99.30	4.43
	$+\mathcal{L}_A$ = ANYRES	**99.42**	**4.02**
128×128	$\mathcal{L}_{base}$	97.53	8.57
	$+\mathcal{L}_P$	98.33	7.25
	$+\mathcal{L}_I$	98.87	5.65
	$+\mathcal{L}_{GAN}^{Local}$	99.36	4.18
	$+\mathcal{L}_A$ = ANYRES	**99.44**	**3.82**

we demonstrate in Figure 6.6 the visual impact of different loss functions. The first baseline experiment is denoted as $\mathcal{L}_{base}$ and considers the method restricted to both, adversarial global loss $\mathcal{L}_{GAN}^{Global}$ and conditional loss $\mathcal{L}_{cond}$, without visual quality optimization. The baseline experiment shows the benefit of using *Squeeze and Excitation* blocks since $\mathcal{L}_{base}$-based results for resolutions 16×16 and 32×32 outperform other methods from the set of ARL-VTF in Table 6.2. In addition, resolutions 64×64 and 128×128 achieve close second best FR scores. However, we note that generated results are rather blurry. By adding additional loss functions to $\mathcal{L}_{base}$, $\mathcal{L}_P$ and $\mathcal{L}_I$, namely the perceptual and identity losses, the scores marginally increase for all resolutions. $\mathcal{L}_{GAN}^{Local}$ improves further. Finally, the proposed ANYRES is able to generate more realistic images with less artefacts at any resolution and with higher facial recognition performances, thus approving the interest of using a combination of above specific loss functions.

6.4.4 Performances with unseen resolutions

Integrated in monitoring systems, CFR systems imply long-range distance acquisition. Such acquisitions involve capturing images at varying distances, *i.e.*, uncontrolled resolutions, as we illustrate in Figure 6.1.

In order to show the versatility of ANYRES addressing a number of resolutions, we select following scale of resolutions: 24×24, 48×48, 96×96 and 112×112. Since the model has not been trained with such resolutions, we denote this experimentation by *unseen resolutions*. Table 6.5 reports accuracies achieved by ANYRES on above resolutions. Results highlight stable performances across all protocols and datasets, thus proving the ability to super-resolve with great face recognition accuracy unseen resolutions.

6.4.5 Towards an end-to-end biometric system

In this Section we combine our TFLD (Chapter 4) with ANYRES, as successive operations, towards designing an end-to-end biometric system.

Related to this experiment, TFLD is targeted to detect faces, as well as specific facial landmarks that are used for image alignment. We note that the process of image alignment follows the same protocol as introduced in Section 4.5.3. ANYRES, on the other hand, translates the aligned image into a synthetic high-resolution visible-like image.

Table 6.6 reports the biometric performances obtained from the proposed experiment on the ARL-VTF dataset. We note that the VIS-TH dataset used in our evaluation and comparison, as presented in Section 6.4.2, does not include facial annotations. Therefore, we have already employed TFLD to preprocess thermal faces in our previous experiments. In contrast to the results presented in Table 6.2, we observe that facial images aligned by TFLD-based facial landmarks achieve higher biometric scores compared to those aligned by the GT annotations, at any given resolution.

To assess the operational feasibility of our solution in real-world scenarios, we conducted experiments to measure the *inference times* of TFLD and ANYRES using both CPU and GPU processors. The results, presented in a Table 6.7, demonstrate that it takes less than a second to *detect* and *translate* a face. This rapid processing time is highly realistic and aligns with the requirements of a practical operation system.

As a result, TFLD is not only effective in a wide range of adverse conditions, but also demonstrates its potential as a robust and accurate facial landmark annotator that is crucial for CFR systems. Therefore, we demonstrated the benefits of using TFLD and ANYRES in tandem, as part of an end-to-end CFR system. Specifically, TFLD has been designed to be robust to unconstrained circumstances such as variations in pose, expression, occlusion, poor image quality and long-range distance. On the other hand, ANYRES is capable of translating spectrum images of any resolution scale, while preserving the identity information, even under pose variations. Together, TFLD and ANYRES provide an accurate and reliable solution, which is especially relevant in the development of an end-to-end biometric system, robust to unconstrained environments.

TABLE 6.5: ANYRES performances on unseen resolutions. Experimental results validate accuracy *w.r.t.* facial recognition, namely by AUC % and EER % scores, as well as visual quality, as reported by SSIM %.

Unseen Res.	**ARL-VTF dataset**			**VIS-TH dataset**			**SF dataset**			**Tufts dataset**		
	AUC (%)	EER (%)	SSIM	AUC (%)	EER (%)	SSIM	AUC (%)	EER (%)	SSIM	AUC (%)	EER (%)	SSIM
24×24	95.43	9.55	64.34	82.84	24.70	53.33	81.74	25.94	66.72	71.00	34.50	39.39
48×48	99.15	4.80	65.30	88.96	18.67	55.75	85.32	22.60	68.44	79.73	27.44	40.43
96×96	99.44	3.92	67.04	89.60	17.93	57.68	88.37	21.28	68.86	80.81	26.39	41.40
112×112	99.44	3.89	67.04	89.62	17.90	57.73	88.51	21.20	68.89	80.90	26.31	40.43

TABLE 6.6: Quantitative comparison of ANYRES on the ARL-VTF dataset when TFLD annotations are used for image alignment. Experimental results validate accuracy *w.r.t.* facial recognition, namely by AUC % and EER % scores.

TFLD + ANYRES-multi Res.	**ARL-VTF dataset** AUC	EER
16×16	91.39	16.70
32×32	98.75	5.80
64×64	99.44	3.98
128×128	99.45	3.65

TABLE 6.7: Average Inference Time of **TFLD** and **ANYRES** on the ARL-VTF dataset, with Nvidia Quadra RTX6000 GPU and Xeon SP Silver CPU.

Inference Time (ms)	GPU	CPU
TFLD	24.2	198.94
ANYRES	5.22	116.67
TFLD + ANYRES	29.42	315.61

6.5 Discussion

6.5.1 Choice of the architecture

ANYRES benefits from a *Pyramid-like* architecture, and we hereby proceed to motivate our choice. Towards addressing SR and CFR, the greatest advantage of pyramid representation is the capability to convert global image features into local features, while condensing representation of the whole image. In such a process, successive levels of the pyramid become a reduced-resolution version of the image, thus relying on multi-scale analysis. ANYRES is built only with one single network and enables the handling of any resolution unlike existing methods such as AxialGAN, which are targeted towards a fixed resolution.

6.5.2 Resolution considerations in ANYRES for CFR

In order to demonstrate the versatility of the method we train ANYRES on 4 resolutions (see Section 6.3.4). This choice stems from constraints related to adversarial training. During the first epochs, the discriminator absorbs many samples of different resolution, then, overpowering the generator would have a negative effect on training. We here note that, being adaptive to the practical CFR scenario by considering 4 scales of resolution is enough and fully reliable for GAN training. While we tested up to an image resolution of 128x128, the method is extendable to higher resolutions. We note that we selected the image resolution of 128x128 for the purpose of comparison of ANYRES with prior works, as well as with the consideration that such resolution is sufficient in the context of CFR. The latter consideration stems from the fact that most industrial and academic facial biometric systems utilize image resolutions, which do not exceed 128x128. For instance, ArcFace [29], which we utilized in

this Chapter for the experimentation, normalizes all faces to an image resolution of 112x112. Additional experiments suggest that ANYRES is capable to super-resolve with remarkable FR accuracy intermediate resolutions (see Section 6.4.4). However, particular attention should be placed on very LR images, where nearly no biometric information is contained, rendering the unveiling of the visible face impossible.

6.5.3 Challenges in handling facial poses in Thermal-to-Visible CFR and the limitations of *Frontalization*

ANYRES has been able to generate faces with different facial poses, thereby enhancing its applicability in real-world and unconstrained situations. However, the challenge with capturing a face in profile has to do with the pose naturally occluding certain facial biometric information, which ultimately affects the performance of the matcher (herein ArcFace) in extracting salient biometric characteristics from the face, as observed in Table 6.3, where the *pose* subset of ARL-VTF was compared *w.r.t.* the baseline subset used throughout the thesis. Despite this, our decision was to synthesize a facial image with the same pose as the captured thermal face instead of pursuing *frontalization*, which involves generating a synthetic frontal face image, while hallucinating information, not present in the input image. The concept of frontalization has been extensively explored in the context of visible spectrum FR systems [145], which in fact benefits from large-scale face image datasets for training. In particular, as highlighted in Chapter 3, this is not the case for thermal data. The limited amount of training data available does not guarantee that the generated frontal face will accurately capture the biometric information contained in the thermal profile face image.

Nevertheless, we explored the performance of ANYRES when adapted to frontalization, *i.e., pose-to-frontal* settings. As expected and depicted in Figure 6.7, global facial appearances are not preserved, or only minimally preserved. This is further confirmed in Table 6.8, where biometric accuracy decreases compared to the process of *pose-to-pose* generation (Table 6.3). However, we draw attention to local facial areas that have been targeted by the local discriminators (see Figure 6.3). These facial regions, namely the eyes, nose and mouth, showcase more consistence to the reality.

This is a significant challenge for thermal-to-visible FR systems, and highlights the importance of further research in this area, targeted to improve the accuracy and reliability of automated biometric identification systems. In this regards, a first attempt to heterogeneous face frontalization has been addressed in [32] via a Domain Agnostic Learning.

6.5.4 Comparative Analysis: Traditional Visible FR *vs.* Thermal-to-Visible CFR

We present a comparative analysis from the ARL-VTF dataset, contrasting traditional visible FR with thermal-to-visible CFR at varying resolutions. Note that, the facial matcher ArcFace [29] has been trained on a large scale visible face

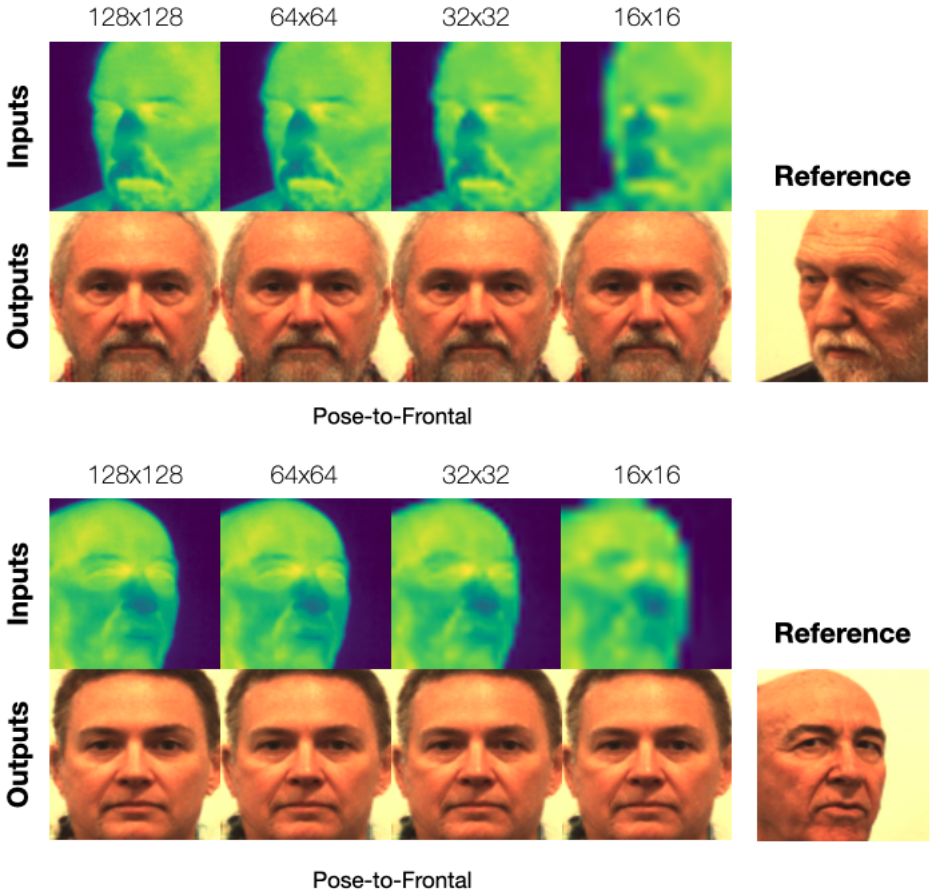


FIGURE 6.7: Qualitative results of ANYRES (pose-to-frontal) on the *pose* subset of the ARL-VTF dataset. We decrease resolution in each row (re-scaled to 128×128).

dataset as discussed in Section 2.3.3, thus enabling to provide more accurate deep characteristic from depicted *visible* faces. Therefore, applying ArcFace directly to thermal face images, as a *direct comparison* approach, is related to the lowest AUC and highest EER scores, as reported from the Table 6.9 (and Table 3.7 *w.r.t.,* other paired thermal-visible face dataset).

While our findings from Table 6.2 showcase higher performance improvement over the direct comparison approach, the CFR method (*i.e.,* ANYRES) produces inferior results, however with a slightly AUC gap in higher resolution, when compared to traditional visible spectrum FR as indicated in Table 6.10. In particular, this analysis stem from the same set of identity, where visible face images are used in probe set and gallery set for the comparison. This leads us to ponder whether having an equivalent amount of thermal facial image data would result in superior performance compared to traditional methods operating with visible spectrum face images.

6.6 Summary

CFR systems necessitate accurate and reliable automated models, able to handle a wide range of resolutions. In this Chapter, we proposed ANYRES, a

TABLE 6.8: Quantitative comparison of pose-to-frontal on the subset *pose* of ARL-VTF dataset. The experimental results validate accuracy *w.r.t.* facial recognition, namely by AUC % and EER % scores.

ANYRES-multi	***pose*-ARL-VTF dataset** *pose-to-frontal*	
Res.	AUC	EER
16×16	73.76	32.19
32×32	79.47	28.30
64×64	80.18	27.82
128×128	80.04	27.86

TABLE 6.9: Quantitative comparison (with using ArcFace FR matcher) on the ARL-VTF dataset when a direct thermal-to-visible cross-spectral face recognition is applied. Experimental results validate accuracy *w.r.t.* facial recognition, namely by AUC % and EER % scores.

Direct Thermal-to-Visible CFR	**ARL-VTF dataset**	
Res.	AUC	EER
16×16	51.30	50.01
32×32	51.71	49.58
64×64	55.05	45.91
128×128	55.16	46.18

TABLE 6.10: Quantitative comparison (with using ArcFace FR matcher) on the ARL-VTF dataset when a traditional face recognition is applied on a set of visible face images. Experimental results validate accuracy *w.r.t.* facial recognition, namely by AUC % and EER % scores.

Traditional Visible FR	**ARL-VTF dataset**	
Res.	AUC	EER
16×16	97.55	7.28
32×32	100	0
64×64	100	0
128×128	100	0

unified generative model that accepts facial images of a wide range of resolutions as input, even under pose variation, and that proceeds to accurately translate such images from one spectrum to another, while ensuring faithful cross-spectral identity. Experiments on four datasets suggest that our proposed ANYRES outperforms state-of-the-art methods. While this is a first step to rendering CFR systems adaptive to real world scenarios, our TFLD combined with ANYRES offers an end-to-end reliable recognition system in unconstrained environments.

Chapter 7

Conclusions and Perspectives

We have advanced the frontier of CFR research, while tackling real-world applications of FR in unconstrained environments. To follow our industrial mandate, we have focused on recognizing individuals by using a *thermal* camera only. It is worth noting that thermal[1] imaging provides the largest appearance gap in terms of image rendering in contrast to the visible spectrum, resulting in a more challenging task of thermal-to-visible CFR scenario. This work is however fully adaptable to other infrared sub-bands.

In practical applications, subjects are not expected to be consistently cooperative, hence we have to design a system that accounts for and is robust to variations of pose, resolution and illumination [104]. Towards designing an operational end-to-end CFR system, we integrated successive new algorithms pertaining to thermal face detection, as well as thermal-to-visible spectrum translation, streamlined to bridge the modality gap. Specifically, firstly faces are detected in thermal images and related accurate facial key-points are extracted for aligning faces. The aligned thermal face image is then translated into a high-resolution visible-like face image. As a final step, this synthetic facial image is fed to any off-the-shelf FR algorithm. We note that, in the context of this CIFRE-thesis, the FR operation is supported by the facial recognition platform (FRP) [38] of Thales, which extracts and compares facial features to the stored gallery template.

Revolving around GANs, we are able to preserve faithful biometric features across *spectra*, as well as *resolutions*, even in diverse range of factors such as long distances acquisition and facial poses. To the best of our knowledge, our approach relying on a unique model handling simultaneously the dual computer vision task of *domain translation* and *super resolution*, has been the first endeavor in this direction.

We proceed to summarize the contributions of this thesis in Section 7.1, whereas limitations and perspectives for future research are outlined in Sections 7.2 and 7.3, respectively.

[1] Thermal denotes here the long-wave infrared (LWIR) band.

7.1 Summary of contributions

In Chapter 2, we presented a comprehensive overview of state-of-art CFR methods, highlighting benefits and limitations. This overview motivated our research, the explored CFR scenario, as well as proposed algorithms aimed at mitigating such limitations. While the NIR spectrum has been extensively explored, with accuracies of associated FR-systems comparable to FR-models in the visible spectrum, other infrared sub-bands such as SWIR and MWIR have been less studied due to their cost and limited availability of training datasets. However, LWIR imaging has shown promising results and is becoming more affordable, rendering it as suitable choice for LWIR-to-VIS FR.

In Chapter 3, we introduced BYDB, a novel dataset developed in an industrial contribution for our research work. Particular efforts were made to mitigate deficiencies of existing datasets, such as including images in an *unconstrained* environment. Considered at the time of writing this dissertation as the largest collection of paired multi-spectral face images, BYDB has contributed to evaluate our models, thereby enhancing accuracy of each component of the pipeline that we proposed towards designing an end-to-end biometric system. Overall, BYDB has been an invaluable resource in our research and has contributed significantly to our understanding of CFR in uncontrolled environments.

In Chapter 4, we presented TFLD, a novel thermal face and landmark detector designed to be robust to adversarial conditions. In particular, TFLD is integrated as a preprocessing operation in the pipeline of an end-to-end CFR system, thus enabling us to localize an individual and his face in the thermal image, and providing accurate facial key points from that face for the purpose of image face alignment. Finally, this aligned thermal face image is fed to the system for the ultimate task of FR. Unlike prior work, TFLD was designed to focus on texture analysis, utilizing regions of interest rather than specific marks for facial landmark prediction. The paradigm of face and landmarks has been shifted from traditional methods to an object detection task. Such strategy allowed us to overcome the scarcity of high-frequency information encountered by neural networks in the context of thermal images. Finally, TFLD accurately predicts facial landmarks in several benchmarks, including real-world images. We have additionally demonstrated the positive impact of face alignment based on TFLD for related CFR.

In Chapter 5, we introduced two models namely LG-GAN and AG-GAN(+), for accurately translating thermal face images into synthetic visible images, while preserving identity. LG-GAN is an explainable generative model that disentangles identity features from other factors, with specific loss functions ensuring that biometric characteristics are learned and preserved. AG-GAN and AG-GAN+ enhance identity representation features with attention modules (trained in supervised or unsupervised manner, respectively), and extend loss functions from LG-GAN to increase identity preservation and good quality

image generation. Our experiments showed that both models, along with customized loss functions, ensured faithful identity preservation across spectra and generated realistic faces in the visible spectrum, improving CFR performances. LG-GAN and AG-GAN(+) also offer critical insights into finding representations that allow for extracting robust biometric features beyond the visible spectrum and for interpreting the generation process from thermal to visible face images. However, both models lack the ability to handle real-world scenarios and were not suitable for unconstrained CFR settings, which led us to the development of the final contribution of the thesis.

In Chapter 6, we made CFR systems adaptable to real-world scenarios using the methodology developed in Chapter 5. Our model, ANYRES, is a unified generative model with extended loss functions focusing on salient facial areas, and uses a pyramid-like encoder-decoder architecture for domain translation and super-resolution. It is designed to handle any (low) resolution of thermal face images as input, even with facial pose, and to generate a synthetic high-resolution visible-like face image, enabling faithful biometric identity preservation across spectrum and resolution. ANYRES achieves state-of-the-art performance on four benchmark datasets and outperforms existing works under any (low) scale of resolutions. While this is a first step to rendering CFR systems adaptive to real world scenarios, our TFLD functions seamlessly with ANYRES, thereby offering an end-to-end reliable recognition system in unconstrained environments.

7.2 Limitations

Despite the significant advances in performance on commonly used datasets, the aforementioned contributions still encounter certain constraints and limitations when implemented in real-world scenarios.

Critical importance of demographic representation for effective deployment. The lack of data, coupled with the associated limited demographic representation in training datasets, significantly impedes the effectiveness of CFR systems. This limitation is critical for the deployment of such systems in real-world scenarios, as performance can be severely impacted across different demographic groups. To address this issue, there is a crucial need for more diverse and representative resources that include individuals from various ethnicity, age, and gender groups. We collected a dataset presented in Chapter 3, BYDB, in an attempt to enhance existing data. While we collected data pertaining to 412 identities, this dataset remains unbalanced w.r.t. distribution of gender (Figure 3.5), age (Figure 3.6) and ethnicity, which is a prerequisite needed for developing unbiased and fair deep learning models. Additionally, in an attempt to mitigate the diversity problem of data, we implemented loss functions aimed at preventing age or gender shift between the input thermal face image and the generated synthetic image. We observed that such information in thermal faces is less sensitive to the facial structure, thus motivating the

implementation of other loss functions than the reconstruction loss (known as L1 loss). In particular, the *attribute loss* comprising of the *age loss* and *gender loss* from Equations 6.12 and 6.13, respectively, has been designed to extract information such as facial attributes from pertinent areas of the face, where heat information has provided the related cue. Although traditional FR has access to a large amount of training data, it still suffers from issues of bias [36]. Solutions such as synthetic data, penalty terms in the loss functions, or re-weighting schemes have been explored to address demographic representation and ensure equal performance across different demographic groups. Nevertheless, increasing the number of resources in terms of quality, quantity, and diversity remains the most effective way to make significant progress for CFR.

Reduced performances compared to traditional visible spectrum FR systems. Thermal-to-visible CFR has found numerous applications in various fields such as surveillance, defense, and law enforcement. The technology enables the identification of individuals in low-light or no-light conditions [97], even from a distance [134], where traditional visible spectrum FR systems fail to provide reliable results. Although thermal cameras operate during night-time or day-time, it is essential to note that the performance of CFR systems falls short when compared to traditional visible spectrum systems under normal lighting and day-time conditions. However, in situations where visible spectrum FR fails, such as night-time surveillance, thermal-to-visible CFR provides a viable solution. Therefore, we can ask the following questions: *If we had access to a multi-spectral dataset of similar size to the visible spectrum for training, would we achieve comparable biometric performance to traditional FR systems?* Additionally, *if we had access to a thermal face dataset as large as the visible imaging dataset, would we achieve similar performance as traditional FR systems?* Despite the undeniable advantages of thermal imaging, *e.g.,* for nocturnal FR, this electromagnetic spectrum provides images with only limited facial information, resulting in poor detail, low contrast, and little biometric information. Hence, in our opinion, this restriction hinders the models' capability to extract distinguishing biometric data for FR. Moreover, the significant modality gap between visible and thermal imaging makes it difficult to match the performance of traditional FR systems. Therefore, a comprehensive and operational product that utilizes both visible and thermal spectra in tandem could provide a highly reliable solution for a wide range of applications. By incorporating both visible and thermal cameras, such a system could perform exceptionally well in low-light or night-time conditions while still maintaining high accuracy in daylight conditions. Moreover, combining the information from both spectra could lead to improved performance in FR tasks by reducing the effects of various environmental factors. In particular, the thermal spectrum could provide robustness against certain types of disguise, such as the use of masks or other face coverings, which have become more prevalent in recent times. Overall, a global FR system equipped with a dual-sensor camera that combines visible and thermal spectra would offer a highly versatile solution, capable of operating across various scenarios and environmental conditions, while maintaining a high level of security.

7.3 Perspectives

CFR has been a thriving area of computer vision for the past decades, enabling the development of numerous real-world applications in surveillance and non-cooperative recognition. While this thesis has contributed a set of methods that advance and progress the field by addressing pertinent challenges related to thermal-to-visible CFR in unconstrained environments, there remain several promising directions for future research in this area. We envision following research paths.

FR beyond the visible spectrum. Recognizing individuals with cameras operating solely in one infrared sub-band remains a major challenge. This is mainly due to intra-class variability and inherently limited texture and detail information. In addition, limited training data (scarce public available datasets) hinders the design of *deep models*, capable of extracting discriminative facial features for direct FR beyond the visible spectrum [40]. Therefore, common practice is to acquire a gallery of facial images serving as reference for comparison in the visible spectrum and store them as standardized facial images. As a consequence, CFR will remain a pertinent direction and related methods will provide sustainable and reliable solutions in this context.

CFR empowered by model explainability. CFR in this thesis has leveraged generative models, namely GANs, for translating thermal face images into synthetic visible-like images, while preserving the identity. Specifically, we focused on *explainability* through the visualization of learnt *identity code* (LG-GAN contribution) and *attention map* (AG-GAN(+) contribution), which offered insights on pertinent biometric features, as well as improvement of images synthesis (impact of different loss functions). One promising direction for future image interspectral generation, in terms of biometric preservation and image quality, would be to empower image translation by *transparency* and *interpretability*. For example, we envision the exploration of spatial attention and self-attention from Transformers for a an improved feature generation control. By providing insights into the inner working of CFR systems, the incorporation of explainability into biometric recognition systems will be essential in the age of deep learning [106].

Next generation of synthesis models. As the field of deep learning continues to evolve, new and innovative methods for generating synthetic images are emerging. While GANs have been popular in recent years, they still face several limitations such as *mode collapse* and *instability* during training. In contrast, models based on the diffusion process, such as the Diffusion Probabilistic Models (DPMs), offer a promising alternative for generating high-quality synthetic images with greater stability and less distortion (identity better preserved). Hence, transitioning from GANs to DPMs opens up exciting new possibilities for generating interspectral images with improved realism and accuracy, rendering them a worthwhile avenue of future research with new scientific keystones.

In this context, a pioneer work has explored DPM proposing a Denoising Diffusion Probabilistic Model [103]. By exploring and developing these new models, we will advance the field of computer vision and create more sophisticated and reliable systems for a wide range of applications related to unconstrained CFR.

Towards typical CFR real world scenarios. Practical scenarios of application induce unconstrained acquisitions at various environmental conditions, as well as uncontrollable effects, thus challenging CFR. In Chapter 6, we have proposed ANYRES, which is able to handle recognition at any range of distance with a unique model, while being robust to variations in pose. Considering larger range of actions and covering additional effects are instrumental for designing a more adaptive in-the-wild CFR system. In the given context, "actions" and "effects" are referring to various changes or conditions that a person's face may undergo in real-world scenarios, such as changes in pose, expression, lighting, occlusion, makeup, and accessories worn. We plan to account for additional challenges such as thermal images afflicted by blurriness, out of focus and atmospheric turbulence. Particular attention will be placed on outdoor environments, during day or night time. Such can affect human appearances in thermal images, depending notably on the weather. Further investigation on learning identity-*invariance* across spectra [39] rather than frontalization (discussed in Section 6.5.3) as a guidance for image generation is our interest, in which combating intra-personal variations, such as difference in pose and facial expression, can improve CFR accuracy.

Robustness against presentation attacks. CFR has proven to be a valuable tool in monitoring systems, particularly in video surveillance where identifying criminals and suspects is critical. However, FR systems have been found to be vulnerable to presentation attacks, where individuals use various attacks to undermine the system. Such attacks have become more sophisticated and easier to carry out, rendering them more efficient. Related vulnerabilities has become a major concern, driven by the availability of high-quality accessories and instruments for attacks, ranging from simple images to expensive silicone and latex full face masks. Most solutions rely on standard visible or NIR cameras [72, 73] to counter these attacks. However, the use of the thermal band of the electromagnetic spectrum can provide a unique advantage in complementing RGB-imagery for fighting fraud [123, 50]. Thermal imaging provides liveness detection and shape determination, which can enhance the robustness of FR systems against presentation attacks. This approach can significantly improve the effectiveness of biometric recognition systems in detecting and preventing fraud.

Extended applications with infrared imaging for biometric recognition beyond face recognition. Infrared sensors have become more advanced and affordable in recent years, rendering them an attractive option for a number of biometric recognition applications beyond face recognition. Such applications have to do with iris recognition and periocular recognition [4], where near-infrared (NIR) imaging is already commonly used. Additionally, thermal

imaging could be utilized for non-intrusive biometric identification in low-light or no-light settings, without the need for harmful illumination that could potentially harm ocular health (unlike NIR lights). While the use of thermal imaging may not be practical for iris recognition due to the interference of eyeglasses (see Figure 6.1), other biometric applications could benefit from the translation of this technology. As infrared sensors continue to evolve, we expect for opportunities of extended biometric recognition beyond the face to mature.

Bibliography

[1] A. F. Abate et al. "2D and 3D face recognition: A survey". In: *Pattern Recognition Letters* 28.14 (2007), pp. 1885 –1906.

[2] M. Abdrakhmanova et al. "Speakingfaces: A large-scale multimodal dataset of voice commands with visual and thermal video streams". In: *Sensors* (2021).

[3] M.Z. Alom et al. "The history began from alexnet: A comprehensive survey on deep learning approaches". In: *arXiv preprint arXiv:1803.01164* (2018).

[4] Fernando Alonso-Fernandez et al. "Periocular Biometrics: A Modality for Unconstrained Scenarios". In: *arXiv preprint arXiv:2212.13792* (2022).

[5] D. Anghelone et al. "Beyond the Visible: A Survey on Cross-spectral Face Recognition". In: *preprint* (2022).

[6] David Anghelone, Sarah Lannes, and Antitza Dantcheva. "ANYRES: Generating High-Resolution visible-face images from Low-Resolution thermal-face images". In: *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2023, pp. 246–251.

[7] David Anghelone et al. "Explainable Thermal to Visible Face Recognition Using Latent-Guided Generative Adversarial Network". In: *IEEE International Conference on Automatic Face & Gesture Recognition, FG 2021, Jodhpur, India, Dec 15-18, 2021*. IEEE, 2021.

[8] David Anghelone et al. "TFLD: Thermal Face and Landmark Detection for Unconstrained Cross-spectral Face Recognition". In: *2022 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE. 2022, pp. 1–9.

[9] L. Assirati et al. "Performing edge detection by difference of gaussians using q-gaussian kernels". In: *Journal of Physics: Conference Series*. Vol. 490. 1. IOP Publishing. 2014, pp. 012–020.

[10] M.K. Bhowmik et al. "Thermal infrared face recognition—a biometric identification technique for robust security system". In: *Reviews, refinements and new ideas in face recognition* 7 (2011).

[11] M. Bihn et al. "Evaluating a Convolutional Neural Network on Short-Wave Infra-Red Images". In: *IEEE Winter Applications of Computer Vision Workshops*. 2018, pp. 18–27.

[12] T. Bourlai and B. Cukic. "Multi-spectral face recognition: Identification of people in difficult environments". In: *IEEE International Conference on Intelligence and Security Informatics*. 2012, pp. 196–201.

[13] T. Bourlai and L.A. Hornak. "Face recognition outside the visible spectrum". In: *Image and Vision Computing* 55 (2016), pp. 14 –17.

[14] T. Bourlai et al. "Cross-Spectral Face Verification in the Short Wave Infrared (SWIR) Band". In: *International Conference on Pattern Recognition*. 2010, pp. 1343–1347.

[15] P. Buddharaju et al. "Physiology-Based Face Recognition in the Thermal Infrared Spectrum". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.4 (2007), pp. 613–626.

[16] K.A. Byrd. "Preview of the newly acquired NVESD-ARL multimodal face database". In: *SPIE*. Vol. 8734. 2013, pp. 8734–8734.

[17] B. Cao et al. "Multi-Margin based Decorrelation Learning for Heterogeneous Face Recognition". In: *International Joint Conference on Artificial Intelligence*. 2019, pp. 680–686.

[18] Z. Cao, N.A. Schmid, and T. Bourlai. "Composite multilobe descriptors for cross-spectral recognition of full and partial face". In: *Optical Engineering* 55.8 (2016), pp. 1 –15.

[19] S. Chatterjee and W.T. Chu. "Thermal Face Recognition Based on Transformation by Residual U-Net and Pixel Shuffle Upsampling". In: *International Conference on Multimedia Modeling*. 2020, pp. 679–689.

[20] C. Chen et al. "Attention-Guided Generative Adversarial Network for Explainable Thermal to Visible Face Recognition". In: *2022 IEEE International Joint Conference on Biometrics (IJCB 2022)*.

[21] Cunjian Chen and Arun Ross. "Matching Thermal to Visible Face Images Using a Semantic-Guided Generative Adversarial Network". In: *IEEE International Conference on Automatic Face & Gesture Recognition*. 2019, pp. 1–8.

[22] J. Chen et al. "Learning mappings for face synthesis from near infrared to visual light images". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 156–163.

[23] Sheng Chen et al. "MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices". In: *Proc. of CCBR*. 2018.

[24] M. Cho et al. "Relational Deep Feature Learning for Heterogeneous Face Recognition". In: *IEEE Transactions on Information Forensics and Security* 16 (2020), 376–388.

[25] Wei-Ta Chu and Yu-Hui Liu. "Thermal facial landmark detection by deep multi-task learning". In: *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2019, pp. 1–6.

[26] WRIGHT STATE Dataset. "https://wsri.wright.edu/". In: URL: https://wsri.wright.edu/.

[27] T. de Freitas Pereira, A. Anjos, and S. Marcel. "Heterogeneous Face Recognition Using Domain Specific Units". In: *IEEE Transactions on Information Forensics and Security* 14.7 (2019), pp. 1803–1816.

[28] J. Deng, Y. Zhou, and S. Zafeiriou. "Marginal loss for deep face recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017, pp. 60–68.

[29] Jiankang Deng et al. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 4690–4699.

[30] Z. Deng et al. "Mutual Component Convolutional Neural Networks for Heterogeneous Face Recognition". In: *IEEE Transactions on Image Processing* 28.6 (2019), pp. 3102–3114.

[31] X. Di et al. "Polarimetric Thermal to Visible Face Verification via Self-Attention Guided Synthesis". In: *International Conference on Biometrics*. IEEE, 2019, pp. 1–8.

[32] Xing Di, Shuowen Hu, and Vishal M Patel. "Heterogeneous Face Frontalization via Domain Agnostic Learning". In: *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. 2021, pp. 01–08. DOI: 10.1109/FG52635.2021.9666962.

[33] Xing Di, He Zhang, and Vishal M Patel. "Polarimetric thermal to visible face verification via attribute preserved synthesis". In: *IEEE International Conference on Biometrics Theory, Applications and Systems*. 2018.

[34] Xing Di et al. "Multi-Scale Thermal to Visible Face Verification via Attribute Guided Synthesis". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3.2 (2021), pp. 266–280.

[35] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *ICLR* (2021).

[36] Pawel Drozdowski et al. "Demographic bias in biometrics: A survey on an emerging challenge". In: *IEEE Transactions on Technology and Society* 1.2 (2020), pp. 89–103.

[37] V. Espinosa-Duró, M. Faundez-Zanuy, and J. Mekyska. "A New Face Database Simultaneously Acquired in Visible, Near-Infrared and Thermal Spectrums". In: *Cognitive Computation* 5 (2013), pp. 119–135.

[38] *Facial Recognition Platform (FRP) of Thales*. https://www.thalesgroup.com/en/markets/digital-identity-and-security/government/biometrics/biometric-software/live-face-identification-system. Accessed: 2023.

[39] Cedric Nimpa Fondje, Shuowen Hu, and Benjamin S Riggan. "Learning Domain and Pose Invariance for Thermal-to-Visible Face Recognition". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* (2022).

[40] Cedric Nimpa Fondje et al. "Cross-domain identification for thermal-to-visible face recognition". In: *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE. 2020, pp. 1–9.

[41] Chaoyou Fu et al. "Dual Variational Generation for Low-Shot Heterogeneous Face Recognition". In: *CoRR* abs/1903.10203 (2019).

[42] Chaoyou Fu et al. "DVG-Face: Dual Variational Generation for Heterogeneous Face Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).

[43] K. Fukushima. "Neocognitron: A hierarchical neural network capable of visual pattern recognition". In: *Neural networks* 1.2 (1988), pp. 119–130.

[44] R.S. Ghiass et al. "Infrared face recognition: A comprehensive review of methodologies and databases". In: *Pattern Recognition* 47.9 (2014), pp. 2807 –2824. ISSN: 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2014.03.015. URL: http://www.sciencedirect.com/science/article/pii/S0031320314001137.

[45] I.J. Goodfellow et al. *Generative Adversarial Networks*. 2014.

[46] Ka. He et al. "Deep residual learning for image recognition". In: *IEEE conference on Computer Vision and Pattern Recognition*. 2016, pp. 770–778.

[47] R. He et al. "Cross-spectral Face Completion for NIR-VIS Heterogeneous Face Recognition". In: *ArXiv* abs/1902.03565 (2019).

[48] R. He et al. "Learning Invariant Deep Representation for NIR-VIS Face Recognition". In: *AAAI Conference on Artificial Intelligence*. 2017, 2000–2006.

[49] R. He et al. "Wasserstein cnn: Learning invariant features for nir-vis face recognition". In: *IEEE transactions on pattern analysis and machine intelligence* 41.7 (2018), pp. 1761–1773.

[50] G. Heusch et al. "Deep Models and Shortwave Infrared Information to Detect Face Presentation Attacks". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* (2020).

[51] J. Hu, L. Shen, and G. Sun. "Squeeze-and-excitation networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[52] S. Hu et al. "A Polarimetric Thermal Database for Face Recognition Research". In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2016.

[53] Shuowen Hu et al. "Heterogeneous Face Recognition: Recent Advances in Infrared-to-Visible Matching". In: *IEEE International Conference on Automatic Face Gesture Recognition*. 2017, pp. 883–890.

[54] W. Hu and H. Hu. "Adversarial Disentanglement Spectrum Variations and Cross-modality Attention Networks for NIR-VIS Face Recognition". In: *IEEE Transactions on Multimedia* (2020), pp. 1–1.

[55] W. Hu and H. Hu. "Discriminant Deep Feature Learning based on joint supervision Loss and Multi-layer Feature Fusion for heterogeneous face recognition". In: *Computer Vision and Image Understanding* 184 (2019), pp. 9–21.

[56] W. Hu and H. Hu. "Disentangled Spectrum Variations Networks for NIR–VIS Face Recognition". In: *IEEE Transactions on Multimedia* 22.5 (2020), pp. 1234–1248.

[57] W. Hu and H. Hu. "Dual Adversarial Disentanglement and Deep Representation Decorrelation for NIR-VIS Face Recognition". In: *IEEE Transactions on Information Forensics and Security* (2020).

[58] W. Hu, H. Hu, and X. Lu. "Heterogeneous Face Recognition Based on Multiple Deep Networks With Scatter Loss and Diversity Combination". In: *IEEE* 7 (2019), pp. 75305–75317.

[59] D Huang, J Sun, and Y Wang. "The BUAA-VisNir face database instructions". In: *School Comput. Sci. Eng., Beihang Univ., Beijing, China, Tech. Rep. IRIP-TR-12-FR-001* 3 (2012), p. 3.

[60] G. Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4700–4708.

[61] Xun Huang and Serge Belongie. "Arbitrary style transfer in real-time with adaptive instance normalization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 1501–1510.

[62] Xun Huang et al. "Multimodal Unsupervised Image-to-image Translation". In: *European Conference on Computer Vision*. 2018.

[63] R. Immidisetti, S. Hu, and V. M. Patel. "Simultaneous face hallucination and translation for thermal to visible face verification using axial-gan". In: *2021 IEEE International Joint Conference on Biometrics (IJCB)*.

[64] Seyed Mehdi Iranmanesh and Nasser M. Nasrabadi. "Attribute-Guided Deep Polarimetric Thermal-to-visible Face Recognition". In: *Proc. of ICB*. 2019.

[65] S.M. Iranmanesh and N.M. Nasrabadi. *Attribute-Guided Deep Polarimetric Thermal-to-visible Face Recognition*. 2019. arXiv: 1907.11980 [cs.CV].

[66] S.M. Iranmanesh et al. "Coupled Generative Adversarial Network for Heterogeneous Face Recognition". In: *Image and Vision Computing* (2019), p. 103861.

[67] S.M. Iranmanesh et al. "Deep cross polarimetric thermal-to-visible face recognition". In: *International Conference on Biometrics*. IEEE. 2018, pp. 166–173.

[68] Phillip Isola et al. "Image-to-Image Translation with Conditional Adversarial Networks". In: *Proc. of CVPR*. 2017.

[69] A.K. Jain, A. Ross, and S. Prabhakar. "An introduction to biometric recognition". In: *IEEE Transactions on circuits and systems for video technology* 14.1 (2004), pp. 4–20.

[70] Anil K. Jain, Debayan Deb, and Joshua J. Engelsma. "Biometrics: Trust, But Verify". In: *IEEE Transactions on Biometrics, Behavior, and Identity Science* 4.3 (2022), pp. 303–323. DOI: 10.1109/TBIOM.2021.3115465.

[71] Anil K Jain and Stan Z Li. *Handbook of face recognition*. Vol. 1. Springer, 2011.

[72] F. Jiang, P. Liu, and X. Zhou. "Multilevel fusing paired visible light and near-infrared spectral images for face anti-spoofing". In: *Pattern Recognition Letters* 128 (2019), pp. 30 –37.

[73] F. Jiang et al. "Face anti-spoofing with generated near-infrared images". In: *Multimedia Tools and Applications* (2020).

[74] J. Johnson, A. Alahi, and L. Fei-Fei. "Perceptual Losses for Real-Time Style Transfer and Super-Resolution". In: *European Conference on Computer Vision*. Ed. by B. Leibe et al. 2016, pp. 694–711.

[75] N. D. Kalka et al. "Cross-spectral face recognition in heterogeneous environments: A case study on matching visible to short-wave infrared imagery". In: *International Joint Conference on Biometrics*. 2011.

[76] A. Kantarcı and H.K. Ekenel. "Thermal to Visible Face Recognition Using Deep Autoencoders". In: *2019 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE. 2019, pp. 1–5.

[77] Jin Keong et al. "Multi-spectral facial landmark detection". In: *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE. 2020, pp. 1–6.

[78] L. Kezebou et al. "TR-GAN: thermal to RGB face synthesis with generative adversarial network for cross-modal face recognition". In: *Mobile Multimedia/Image Processing, Security, and Applications*. Vol. 11399. SPIE, 2020, pp. 158 –168.

[79] Junho Kim et al. "U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation". In: *Proc. of ICLR*. 2020.

[80] Davis E King. "Dlib-ml: A machine learning toolkit". In: *The Journal of Machine Learning Research* 10 (2009), pp. 1755–1758.

[81] B.F. Klare and A.K. Jain. "Heterogeneous face recognition using kernel prototype similarities". In: *IEEE transactions on pattern analysis and machine intelligence* 35.6 (2013), pp. 1410–1422.

[82] S.G. Kong et al. "Recent advances in visual and infrared face recognition review". In: *Computer Vision and Image Understanding* 97.1 (2005), pp. 103 –135. ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2004.04.001. URL: http://www.sciencedirect.com/science/article/pii/S1077314204000451.

[83] M. Kopaczka et al. "A thermal infrared face database with facial landmarks and emotion labels". In: *IEEE Transactions on Instrumentation and Measurement* (2018).

[84] M. Kopaczka et al. "A Thermal Infrared Face Database With Facial Landmarks and Emotion Labels". In: *IEEE Transactions on Instrumentation and Measurement* 68.5 (2019), pp. 1389–1401.

[85] M. Krišto and M. Ivasic-Kos. "An overview of thermal face recognition methods". In: *International Convention on Information and Communication Technology, Electronics and Microelectronics*. 2018, pp. 1098–1103.

[86] Askat Kuzdeuov et al. "SF-TL54: A Thermal Facial Landmark Dataset with Visual Pairs". In: *2022 IEEE/SICE International Symposium on System Integration (SII)*. IEEE. 2022, pp. 748–753.

[87] Askat Kuzdeuov et al. "TFW: Annotated Thermal Faces in the Wild Dataset". In: (2022).

[88] Y. LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[89] C. Ledig et al. "Photo-realistic single image super-resolution using a generative adversarial network". In: *IEEE conference on computer vision and pattern recognition*. 2017.

[90] Cheng-Han Lee et al. "MaskGAN: Towards Diverse and Interactive Facial Image Manipulation". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2020.

[91] J. Lezama, Q. Qiu, and G. Sapiro. "Not Afraid of the Dark: NIR-VIS Face Recognition via Cross-Spectral Hallucination and Low-Rank Embedding". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

[92] S. Z. Li and A. K. Jain, eds. *Handbook of Face Recognition*. Springer London, 2011. DOI: 10.1007/978-0-85729-932-1.

[93] Stan Z. Li, Zhen Lei, and Meng Ao. "The HFB Face Database for Heterogeneous Face Biometrics research". In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 2009, pp. 1–8. DOI: 10.1109/CVPRW.2009.5204149.

[94] Stan Z. Li et al. "The CASIA NIR-VIS 2.0 Face Database". In: *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2013, pp. 348–353. DOI: 10.1109/CVPRW.2013.59.

[95] W. Liu et al. *SphereFace: Deep Hypersphere Embedding for Face Recognition*. 2017. arXiv: 1704.08063 [cs.CV].

[96] X. Liu et al. "Transferring deep representation for NIR-VIS heterogeneous face recognition". In: 2016, pp. 1–8.

[97] Hyunju Maeng et al. "Nighttime face recognition at long distance: Cross-distance and cross-spectral matching". In: *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part II 11*. Springer. 2013, pp. 708–721.

[98] K Mallat and J-L Dugelay. "A benchmark database of visible and thermal paired face images across multiple variations". In: *International Conference of the Biometrics Special Interest Group, BIOSIG 2018, Darmstadt, Germany, September*.

[99] K. Mallat and J-L. Dugelay. "Facial landmark detection on thermal data via fully annotated visible-to-thermal data synthesis". In: *IEEE International Joint Conference on Biometrics (IJCB)*. IEEE. 2020.

[100] K. Mallat et al. "Cross-spectrum thermal to visible face recognition based on cascaded image synthesis". In: *International Conference on Biometrics*. 2019, pp. 1–8.

[101] I. Masi et al. "Deep Face Recognition: A Survey". In: *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. 2018, pp. 471–478.

[102] R. Munir and R.A. Khan. "An extensive review on spectral imaging in biometric systems: Challenges & advancements". In: *Journal of Visual Communication and Image Representation* 65 (2019).

[103] Nithin Gopalakrishnan Nair and Vishal M. Patel. "T2V-DDPM: Thermal to Visible Face Translation using Denoising Diffusion Probabilistic Models". In: *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. 2023, pp. 1–7. DOI: 10.1109/FG57933.2023.10042661.

[104] Chrisopher B Nalty et al. "A Brief Survey on Person Recognition at a Distance". In: *arXiv preprint arXiv:2212.08969* (2022).

[105] N. Narang, T. Bourlai, and L.A. Hornak. "Can we match ultraviolet face images against their visible counterparts?" In: *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XXI*. Vol. 9472. International Society for Optics and Photonics. 2015, 94721Q.

[106] Pedro C Neto et al. "Explainable biometrics in the age of deep learning". In: *arXiv preprint arXiv:2208.09500* (2022).

[107] University of Notre Dame. "University of Notre Dame biometric data set collection C". In: *https://cvrl.nd.edu/projects/data/nd-2006-data-set* (Last accessed 2012).

[108] University of Notre Dame. "University of Notre Dame biometric data set collection C". In: *https://cvrl.nd.edu/projects/data/nd-2006-data-set* (Last accessed 2012).

[109] K. Panetta et al. "A comprehensive database for benchmarking imaging systems". In: *IEEE transactions on pattern analysis and machine intelligence* (2018).

[110] K. Panetta et al. "A Comprehensive Database for Benchmarking Imaging Systems". In: *IEEE TPAMI* 42.3 (2020), pp. 509–520.

[111] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. "Deep Face Recognition". In: (2015).

[112] N. Peri et al. "A Synthesis-Based Approach for Thermal-to-Visible Face Verification". In: *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition*.

[113] Neehar Peri et al. "A Synthesis-Based Approach for Thermal-to-Visible Face Verification". In: *Proc. of FG*. 2021.

[114] D. Poster et al. "A Large-Scale, Time-Synchronized Visible and Thermal Face Dataset". In: *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2021, pp. 1559–1568.

[115] D. Poster et al. "An Examination of Deep-Learning Based Landmark Detection Methods on Thermal Face Imagery". In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2019, pp. 980–987.

[116] D. Poster et al. "Visible-to-Thermal Transfer Learning for Facial Landmark Detection". In: *IEEE Access* (2021).

[117] Domenick Poster et al. "A Large-Scale, Time-Synchronized Visible and Thermal Face Dataset". In: *IEEE Winter Conference on Applications of Computer Vision*. 2021, pp. 1559–1568.

[118] C. Reale et al. "Seeing the Forest from the Trees: A Holistic Approach to Near-Infrared Heterogeneous Face Recognition". In: 2016, pp. 320–328.

[119] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *International Conference on Medical Image Computing and Computer Assisted Intervention*. 2015, pp. 234–241.

[120] M. Sarfraz and R. Stiefelhagen. "Deep Perceptual Mapping for Cross-Modal Face Recognition". In: *International Journal of Computer Vision* (Jan. 2017).

[121] S. I. Serengil and A. Ozpinar. "HyperExtended LightFace: A Facial Attribute Analysis Framework". In: *2021 IEEE International Conference on Engineering and Emerging Technologies*.

[122] Karen Simonyan and Andrew Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *International Conference on Learning Representations*. 2015.

[123] L. Spinoulas et al. "Multispectral Biometrics System Framework: Application to Presentation Attack Detection". In: *IEEE Sensors Journal* 21.13 (2021), pp. 15022–15041. DOI: 10.1109/JSEN.2021.3074406.

[124] Valeriya Strizhkova et al. "Emotion Editing in Head Reenactment Videos using Latent Space Manipulation". In: *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. 2021, pp. 1–8. DOI: 10.1109/FG52635.2021.9667059.

[125] Y. Sun et al. "Deep learning face representation by joint identification-verification". In: *Advances in neural information processing systems*. 2014, pp. 1988–1996.

[126] Yi Sun, Xiaogang Wang, and Xiaoou Tang. "Deep learning face representation from predicting 10,000 classes". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1891–1898.

[127] C. Szegedy et al. "Going deeper with convolutions". In: *IEEE conference on Computer Vision and Pattern Recognition*. 2015, pp. 1–9.

[128] C. Szegedy et al. "Rethinking the inception architecture for computer vision". In: *Conference on Computer Vision and Pattern Recognition*. 2016, pp. 2818–2826.

[129] Y. Taigman et al. "Deepface: Closing the gap to human-level performance in face verification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1701–1708.

[130] Hao Tang et al. "AttentionGAN: Unpaired Image-to-Image Translation using Attention-Guided Generative Adversarial Networks". In: *IEEE Trans. on NNLS* (2021).

[131] WVHTC Foundation Technologies Group. "PRE-TINDERS and TINDERS Face Database". In: (2011). URL: http://www.wvhtf.org/departments/advanced_tech/projects/tinders.asp..

[132] University of Tennessee. "IRIS thermal/visible face database." In: *http://www.cse.ohio-state.edu/otcbvs-bench/* (2012).

[133] *THALES group*. https://www.thalesgroup.com/en. Accessed: 2023.

[134] Kien Nguyen Thanh et al. "The State of Aerial Surveillance: A Survey." In: *CoRR* (2022).

[135] H. Wang et al. "CosFace: Large Margin Cosine Loss for Deep Face Recognition". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[136] Mei Wang and Weihong Deng. "Deep face recognition: A survey". In: *Neurocomputing* (2020).

[137] Z. Wang, Z. Chen, and F. Wu. "Thermal to Visible Facial Image Translation Using Generative Adversarial Networks". In: *IEEE Signal Processing Letters* 25.8 (2018), pp. 1161–1165.

[138] K. Weiss, T.M. Khoshgoftaar, and D. Wang. "A survey of transfer learning". In: *Journal of Big data* 3.1 (2016), p. 9.

[139] Y. Wen et al. "A discriminative feature learning approach for deep face recognition". In: *European conference on computer vision*. Springer. 2016, pp. 499–515.

[140] L.B. Wolff, D.A. Socolinsky, and C.K. Eveland. "Face recognition in the thermal infrared". In: *Computer Vision Beyond the Visible Spectrum*. 2005, pp. 167–191.

[141] X. Wu et al. "A light cnn for deep face representation with noisy labels". In: *IEEE Transactions on Information Forensics and Security* 13.11 (2018), pp. 2884–2896.

[142] X. Wu et al. "Coupled Deep Learning for Heterogeneous Face Recognition". In: 2018.

[143] L. Yang et al. "Hifacegan: Face renovation via collaborative suppression and replenishment". In: *ACM International Conference on Multimedia*. 2020.

[144] D. Yi et al. "Learning Face Representation from Scratch". In: *ArXiv* abs/1411.7923 (2014).

[145] Y. Yin et al. "Dual-attention GAN for large-pose face frontalization". In: *2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020)*. IEEE. 2020, pp. 249–256.

[146] Aijing Yu et al. "LAMP-HQ: A large-scale multi-pose high-quality database and benchmark for NIR-VIS face recognition". In: *International Journal of Computer Vision* 129.5 (2021), pp. 1467–1483.

[147] He Zhang et al. "Generative adversarial network-based synthesis of visible faces from polarimetrie thermal faces". In: *IEEE International Joint Conference on Biometrics*. 2017, pp. 100–107.

[148] He Zhang et al. "Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks". In: *International Journal of Computer Vision* 127.6 (2019), pp. 845–862.

[149] K. Zhang et al. "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks". In: *IEEE Signal Processing Letters* 23.10 (2016), pp. 1499–1503.

[150] T. Zhang et al. "TV-GAN: Generative Adversarial Network Based Thermal to Visible Face Recognition". In: *2018 International Conference on Biometrics*. 2018, pp. 174–181.

[151] T. Zhang et al. "TV-GAN: Generative Adversarial Network Based Thermal to Visible Face Recognition". In: *International Conference on Biometrics*. 2018, pp. 174–181.

[152] Guoying Zhao et al. "Facial expression recognition from near-infrared videos". In: *Image and vision computing* 29.9 (2011), pp. 607–619.

[153] W. Zhao et al. "Face recognition: A literature survey". In: *ACM Computing Surveys* 35.4 (2003), pp. 399–458.

[154] Bolei Zhou et al. "Learning deep features for discriminative localization". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2921–2929.

[155] J.Y. Zhu et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks". In: *IEEE international conference on computer vision*. 2017, pp. 2223–2232.

[156] Joanne Zwinkels. "Light, electromagnetic spectrum". In: *Encyclopedia of Color Science and Technology* 8071 (2015), pp. 1–8.