Advancing Beyond People Recognition in Facial Image Processing

Nélida Mirabet-Herranz

Advancing Beyond People Recognition in Facial Image Processing

Dissertation

submitted to

Sorbonne Université

in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Author:

Nélida Mirabet-Herranz

Publicly defended on 25/06/24 before a committee composed of:

President/Examiner	**Prof. Amine Nait-Ali**	Université Paris-Est Créteil, FR
Examiner/Reviewer	**Prof. Maria de Marsico**	Sapienza Università di Roma, IT
Examiner/Reviewer	**Prof. Julian Fierrez**	Universidad Autónoma de Madrid, ES
Examiner	**Prof. Maria A. Zuluaga**	EURECOM, FR
Examiner	**Mr. Fabien Aili**	Docaposte Biometrics Lab, FR
Co-supervisor	**Prof. Chiara Galdi**	EURECOM, FR
Thesis Director	**Prof. Jean-Luc Dugelay**	EURECOM, FR

The research has been conducted in the Digital Security Department of EURECOM (Sophia Antipolis, FR) from October 2020 to March 2024.

Acknowledgements

Six years have passed since the first day I set foot in Eurecom. I came to France for my Master's studies with hundreds of insecurities and thousands of dreams, but never in a million years could I have imagined what an incredible journey this adventure would turn out to be.

On the 24th of February 2020, I was back in Valencia after the final exam period of my Master's at Eurecom when I received an email from Prof. Jean-Luc Dugelay. The subject of that email was "Ph.D.?" — self-explanatory. That email marked the beginning of what would become a lifelong achievement. For the trust he placed in me, the first person I want to thank along these lines is Prof. Jean-Luc. During our four-year collaboration as student and supervisor, he provided invaluable ideas and guidance that enabled the extensive work discussed in this thesis. Secondly, I would like to extend my deepest gratitude to my co-supervisor, Prof. Chiara Galdi. After a not-so-prolific first year, her advice and expertise in computer vision and biometrics were a source of inspiration and pushed my work to be published in renowned conferences and journals in our field. I am also grateful to the reviewers and jury members — Maria De Marsico, Julian Fierrez, Amine Nait-Ali, Maria A. Zuluaga, and Fabien Aili — for their valuable comments and insightful questions during the defense and on this manuscript.

Though it was not an easy journey, now that it has reached its conclusion, I cannot deny that I will miss these years. Since 2020, many new faces have arrived, and many old friends have left, reaffirming the ever-changing nature of the south of France. As I write these lines, I am sitting in my beloved office, 373, which, after four years, I still need to check for the number on the door. Following the inverse-distance criterion, I will start by mentioning Sahar, with whom I shared not just an office but also ideas, laughs, worries, advice, and hundreds of students' labs to correct. Within my Eurecom circle, I want to cherish the moments inside and outside Eurecom that I shared with Alexandre, Elisa—friend, flatmate, adventure partner, and series advisor, Elyssa, Federico, Francesco, Mira, Nino—our aperos with jalapeños will be missed, and Sameer. Beyond my comfort zone at the university, many

others have contributed in their own special way to making this journey unforgettable: Emma, Florencia, Hamere, Iulia, Marine—thanks to whom I can now say I speak a little French—Nico, Pietro, Sarah, Simone, Sylwia, Thalyssa, and Zineb. Special thanks to my friends back in Spain: Irene, Julia, Maria, Marta, and Teresa, who always welcomed me with open arms, no matter the distance.

Being apart from my family has been the most heartbreaking and challenging part of my Ph.D. While it is not common practice in research, I have proudly included both of my surnames in my publications, as each side of my family has taught me the values that have led me to my career achievements. I can say without hesitation that without my grandparents Nélida, Florentino, and Julia, I would not be the researcher and the person I am today. I am thankful to all my aunts and uncles—Aselina, Gustavo, Cristina, Ronaldo, Miguel, Gloria, and Julio—who filled my childhood with happy memories and offered encouraging words when it was hard to come back to France after the Christmas holidays. A special place in my heart is reserved for my cousins María— who has accompanied me in life since year one— Irene, César and Valeria.

I also want to express my deep admiration for my parents, Yeni and Juan, whose successful careers never stopped them from being present every day of my life. The gratitude I feel for you cannot be fully expressed in words, as you have comforted and lifted me up more times than I can count during this Ph.D. I also want to share how much love I have for my sister Julia, the funniest person I know. The moments we shared while living together this past year will always be cherished in my heart.

Lastly, there is one person who deserves acknowledgment further than those lines can hold. I want to thank Eugenio, my partner, for his unwavering support and patience. We have walked this path side to side through ups and downs and there was no better ending to the journey than obtaining our Ph.D. in consecutive days. There is nobody else I would have shared these years with.

Nélida Mirabet-Herranz
Eurecom, 20 September 2024

Abstract

Human faces encode a vast amount of information, including distinctive features of an individual and demographic characteristics such as a person's age, gender, and weight. Such information is referred to as soft biometrics, which comprises physical, behavioral, or adhered human characteristics classifiable into predefined human-compliant categories. Additionally, some descriptors, like heart rate, fall into the category of the so-called hidden biometrics, metrics of human physiological activities invisible to the naked eye that can serve to assess a person's health status.

The goal of this thesis is to explore the estimation of biometric traits, namely gender, age, weight, and heart rate from facial visuals. In particular, this manuscript includes contributions on improving deep learning models for automatic and contactless estimation of these traits and seeks to deepen the understanding of the key role that widespread practices of social media uploading and filtering of visuals play in these models and their final prediction.

In the first part of the thesis, we define optimal pipelines and architectures for different facial processing tasks. As a first step, our work focuses on extensively improving state-of-the-art networks for weight and heart rate estimation in the visible spectrum. However, visuals in this spectrum gather limited information. Consequently, a promising approach for soft biometrics estimation is presented, where thermal imaging technology is introduced as both an alternative and complement for soft biometric extraction. New advancements in the facial processing field are made, as in this part, a unique face embedding for multi-biometric estimation is proposed in both the visible and thermal domains. In the second part of the thesis, we study the challenge posed by face beautification through social media filters to face-based deep learning architectures, considering the widespread uploading and processing of face visuals on social media platforms. We assess and extensively discuss the impact of such techniques on different facial processing models, including face recognition, deepfake detection, and the estimation of soft and hidden biometrics.

Résumé

Le visage humain contient une grande quantité d'informations, notamment les traits distinctifs d'un individu et les caractéristiques démographiques telles que l'âge, le sexe et le poids d'une personne. Ces informations sont appelées biométrie douce. Elle comprend des caractéristiques humaines physiques, comportementales ou liées à l'adhésion, classables dans des catégories prédéfinies conformes à l'être humain. En outre, certains descripteurs, comme la fréquence cardiaque, entrent dans la catégorie de ce que l'on appelle la biométrie cachée. C'est-à-dire, les mesures des activités physiologiques humaines invisibles à l'œil nu qui peuvent servir à évaluer l'état de santé d'une personne.

L'objectif de cette thèse est d'explorer l'estimation des traits biométriques tel que le sexe, l'âge, le poids et la fréquence cardiaque à partir des images faciales. En particulier, ce manuscrit comprend des contributions sur l'amélioration des modèles d'apprentissage profond pour l'estimation automatique et sans contact de ces traits et cherche à approfondir la compréhension du rôle clé que les pratiques répandues de téléchargement des médias sociaux et de filtrage des visuels jouent dans ces modèles et leur prédiction finale.

Dans la première partie de cette thèse, nous définissons des pipelines et des architectures optimales pour différentes tâches de traitement facial. Tout d'abord, notre travail se concentre sur l'amélioration des modèles pour l'estimation du poids et de la fréquence cardiaque dans le spectre visible. Cependant, les images dans ce spectre recueillent des informations limitées. Par conséquent, nous présentons une approche prometteuse pour l'estimation biométrique douce où la technologie d'imagerie thermique est introduite à la fois comme une alternative et un complément pour cette extraction. Dans la deuxième partie de cette thèse, nous étudions le défi posé par l'embellissement des visages à travers les filtres des médias sociaux aux architectures d'apprentissage profond, en prenant en compte le téléchargement et le traitement généralisés des visages sur les plateformes de médias sociaux. Enfin, nous évaluons et discutons en détail l'impact de ces techniques sur différents modèles, y compris la reconnaissance des visages, la détection des videos hyper-truqués et l'estimation des biométries douces et cachées.

Contents

List of Figures

List of Tables

List of Abbreviations

A	Ambient light conditions
AI	Artificial Intelligence
AR	Augmented Reality
ARF	Augmented Reality Filters
AUs	Action Units
B	Blue
BP	Blood Pressure
bpm	Beats per minute
CAF	Color Adjustment Filters
Sc	Snapchat
CFR	Cross Face Recognition
3D-CNN	Three Dimensional Convolutional Neural Network
CNN	Convolutional Neural Network
DDPM	Denoising Diffusion Probabilistic Models
DET	Detection Error Trade-off
DEX	Deep EXpectation for age estimation
DF	Deepfake method
F	Female
FF++	Face Forensics Dataset
F2F	Face2Face method
FFMF	Facial Feature Modification Filters
FMR	False Match Rate
FNMR	False Non Match Rate
FNR	False Negative Rate
fps	Frames per second
FR	Face Recognition
FS	FaceSwap method
FSh	FaceShifter method
G	Green
GAN	Generative Adversarial Networks

HR	Heart Rate
hl	Hidden Layers
IARFF	Immersive Augmented Reality Face Filter
Ig	Instagram
kg	Kilograms
LSTM	Long Short Term Memory
M	Male
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
Me	Mean error
MSE	Mean Squared Error
MTCNN	Multi-Cascade Convolutional Neural Network
μ	Mean
N	Normal conditions
NT	NeuralTexture method
O	Occlusion variation
PPG	Photoplethysmography
R	Red
ResNet50	Residual Neural Network with 50 layers
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
ROI	Region of Interest
RR	Respiratory Rate
SF	Smoothing Filter
SNs	Social Networks
StD	Standard Deviation
SVR	Support Vector Regression
TCN	Temporal Convolutional Networks
TL	Transfer Learning
TH	Thermal
VGG	Visual Geometry Group
VIS	Visible
Visuals	Images and videos

Chapter 1

Introduction

LOOKING at each other plays an important role in human interaction since glancing at others' faces is the most common way to recognize colleagues, childhood friends, and loved ones. When describing someone whose identity is unknown, various traits, including gender and approximate age, are used for the description while in different scenarios, such as shopping and medical emergencies where a scale is unavailable, professionals often visually estimate the weight of the target subject. Furthermore, looking at other people's facial expressions allows the observer to understand if a person's heart rate is high by examining their breathing and provides more subjective information such as their emotional state. We can affirm that human faces encode a vast amount of information. As we often say, one look is worth a thousand words[1].

All this information holds significance in our daily lives. The key component of our identification systems, namely passports and ID cards, is our face picture. Additional descriptors such as gender and age are commonly displayed beside the picture and in certain countries, like France and Italy, the height of the person is also documented. Moreover, in the digital age, technology users regularly track their health parameters. The utilization of smartwatches for Heart Rate (HR) monitoring has joined the longstanding practice of weight control using a scale and checking blood pressure with an arm cuff. In this thesis, the focus is placed on the unification of source data for estimating all the above-mentioned parameters. Facial visuals, the term referring to face images and videos, are analyzed in depth. As the human brain can extract various pieces of information from facial appearance, the objective is to employ data gathered on different spectra and Artificial Intelligence (AI) models to automatically obtain the same insights and beyond, since cameras can capture information invisible to the naked eye.

[1]Attributed to Frederick R. Barnard in *Printers Ink*, Dec. 1921, pp. 96-97.

In today's world of the internet, a challenge arises since face visuals as input data are not assured to be a robust information source. Social media sites have become a meaningful part of our daily existence, where users aim to create and share visually appealing content. The latest technology offered on these platforms to enhance visuals is filters. The most popular type is beautification filters, applied by users to make their content more engaging by conforming their features to beauty standards. Since these filters distort the knowledge encoded in faces, the information obtained using beautified data may now be compromised. Consequently, the impact of those modifications is rigorously quantified in this thesis.

Formally, the goal of this Ph.D. dissertation is the investigation of new techniques for face analysis and parameter extraction beyond established systems such as Face Recognition (FR). The challenges posed by the widespread culture of social media filter usage on these models are extensively analyzed and measured. The following lines will present the reader with a short story inspired by Jack Clark's Tech Tales in his Import AI Newsletter[2]. It is designed to illustrate the application of facial analysis technologies in a (not-so-distant) future. Similar to all aspects of life, this prospect is neither utopian nor dystopian. As scientists, we ought to persist in pushing the boundaries of AI without ignoring the ethical considerations of our work.

[Cannes, 2055]

Walking along the main street during peak hours is never a peaceful activity. Distracted pedestrians chuckle at invisible jokes they see in their smart glasses, while families have to walk in line to avoid colliding with electric car plugs integrated alongside each parking space in the city. The holograms exhibited in the shops continuously shift according to the people gazing at the storefront.

When I stop in front of my favorite boutique, it takes less than 3 seconds to analyze my gender and age to display clothes tailored to my preference. Of course, the featured sizes look correct as they were determined based on my height and weight, simultaneously estimated. Live face recognition in public places was officially forbidden by the EU in 2023, but real-time user profiling using face information is allowed, as in 2044 the textile lobbies pressured the EU to adapt the GDPR policies.

I continue towards my destination when I receive a reminder of my doctor's appointment. I access the phone's camera, disable all beauty filters, capture a selfie, and send it via the app. Before reaching the bar at the end

[2]https://jack-clark.net/

of the street, the screen lights up with the response from the physician. My parameters are stable, but he recommends maintaining the weekly medication to control my blood pressure. Efficient as always.

Satisfied, I step into the bar, spotting my friends at our usual table. After all, there's nothing quite like meeting in person.

1.1 Potential of Facial Data

Human face images encode different biometric information whose estimation is of great utility in security and health applications, including but not limited to people identification, health assessment and psychological status evaluation. Face processing represents a well-established research field, actively investigated for over two decades. Face is considered as a ***hard*** biometric trait, implying that face analysis technologies can deduce explicit information, such as a person's identity, from it. It stands as one of the most crucial and frequently utilized human biometric traits. Substantial progress has been achieved, leading to the development of successful applications on computers and mobile devices. Indeed, automatic face recognition consistently ranks among the most actively researched areas in computer vision [1].

Soft biometric traits are supplementary information not inherently distinctive enough for standalone use in face-recognition tasks. This implies that soft biometrics alone cannot unequivocally authenticate a person due to their lack of distinctiveness and permanence. However, they do provide additional insights into the subject, including gender or ethnicity [2], characteristics traditionally employed in video surveillance and security scenarios. Beyond the scope of people identification a big amount of information belonging to an individual has been proved to be embedded in face visuals. These parameters include, but are not limited to, gender, age, height, and weight. Some of those soft biometric traits, also serve as health indicators for assessing an individual's overall health status [3].

Moreover, facial processing from visual content has gained a lot of attention in the past years as it allows for non-invasive contactless monitoring of a subject's health, useful in numerous potential applications. Nowadays, there is a global trend to monitor eHealth parameters without the use of physical devices enabling their estimation in at-risk situations such as medical emergencies and road accidents besides at-home daily monitoring and telehealth. Within the medical domain, the term ***hidden*** biometrics (concerning human vision) refers to specific methods used to measure parameters extracted from medical data thus enabling the use of biosignals for tasks such as individual identification or verification [4]. Parameters including Heart Rate, Saturation

of peripheral oxygen, or Blood Pressure (BP) can be inferred from face images or videos for people's daily quality of life assessment in addition to giving complementary information for a person's identification. These parameters have played crucial roles in media security and forensics, often being referred to in the literature as microsignals [5].

1.1.1 Problem Formalization

Previous work has formalized deep learning approaches and methodologies such as Face Recognition [6] and transfer learning [7]. In this thesis, a formalization of AI-based facial processing is proposed.

Let $\mathcal{D}$ be an electromagnetic spectral domain composed of a d-dimensional feature space $\mathcal{X} \subset \mathbb{R}^d$ with marginal distribution $\mathbb{P}(\mathcal{X})$ and a label space $\mathcal{Y} \subset \mathbb{N}$.

Given a n-face dataset $X = \{x_i\}_{i=1}^n$, where $x_i \in \mathcal{X}$ and their corresponding n-value for each k-biometric trait $Y^k = \{y_j^k\}_{j=1}^n$ where $y_j^k \in \mathcal{Y}$ and $k = 1, ..., m$ with $n, m \in \mathbb{N}$. Then a *Face Processing Task* is defined as a parametric function $\mathfrak{F}_{k,\Theta}$ described by the deep learning model parameters Θ where

$$\begin{aligned} \mathfrak{F}_{k,\Theta} \colon X \times Y^k &\longrightarrow [0,1] \\ (x_i, y_j^k) &\longmapsto \mathbb{P}(Y = y_j^k | X = x_i, \Theta) \end{aligned}$$

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

Thus, any facial processing model $\mathfrak{F}_{k,\Theta}$ aims to learn the optimal parameters Θ so that the probability of correctly estimating the trait k for all n identities is 1.

1.1.2 Research Questions

The main contributions of the presented work lie in the investigation of novel facial processing models beyond established techniques as well as the evaluation of the impact of the current trend of digital beautification through social media filters in biometrics. Several experiments are presented and the results are analyzed in depth, reporting different metrics to be true to international standards and at the same time allowing comparison with previous work.

Formally, this work seeks to address the following Research Questions (RQ):

- **RQ1:** What are the most suitable deep learning architectures and training methodologies for various facial processing tasks?

- **RQ2:** Are there benefits to extracting soft biometrics from thermal rather than visible data?
- **RQ3:** Do beautification filters pose a threat to the integrity of our existing biometric systems?

1.2 Main Contributions and Thesis Structure

The manuscript begins with an introductory part, which comprises this chapter and the following one. In **Chapter 1** we have introduced the thesis and we have presented the research questions that motivated this work. In **Chapter 2**, we provide a technical literature review on advancements in facial processing techniques, with a particular emphasis on soft and hidden face biometrics. We present related works on the impact of beautification in biometrics. In addition, we position our contributions in relation to the current state-of-the-art.

Afterward, the dissertation is divided into three parts. The first part (Part I) investigates optimal methodologies for soft biometric trait estimation. This Part comprises **Chapters 3, 4, 5**. The second one (Part II) analyzes the threat posed by the current trend of digital beautification through social media filters to different biometric models. This Part is composed by **Chapters 6, 7**. Finally, Part III closes the thesis with **Chapter 8** containing conclusions, future work, and the **Author's publications**. The contributions made by the research included in this thesis are as follows:

Chapter 3

In **Chapter 3**, the first contributions of this thesis are presented, focusing on developing existing models for soft and hidden biometric traits crucial for health assessment: heart rate and weight. Despite extensive research on facial identification, gender, and age prediction, with established optimal models, a performance gap remains for weight and heart rate estimation. The chapter addresses this gap by defining optimal training methodologies and by examining various data preprocessing methods for both networks.

Part of the work presented in this chapter was published in:

- **N. Mirabet-Herranz**, K. Mallat, J-L. Dugelay, **"Deep Learning for Remote Heart Rate Estimation: A Reproducible and Optimal State-of-the-Art Framework"** in *International Conference on Pattern Recognition (ICPR)*, Montreal, Canada, August 2022.

- **N. Mirabet-Herranz**, K. Mallat, J-L. Dugelay, **"New Insights on Weight Estimation from Face Images"** in *IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, Hawaii, USA, January 2023.

Chapter 4

Chapter 4 outlines the efforts in creating a dual visible-thermal, paired-by-design face dataset encompassing annotations for over 22 traits, including identity and various soft and hidden biometrics. The meticulous design of this dataset not only contributes to the advancement of research in the field but also serves as the fundamental basis for other works presented in **Chapters 4, 5**. The chapter includes a comprehensive review of existing thermal datasets, along with details about the design and discussion of the collection protocol for this dataset. Furthermore, it reports the results of experiments conducted to validate thermal imagery effectiveness in estimating different soft biometrics traits.

Part of the work presented in this chapter was published in:

- **N. Mirabet-Herranz**, J-L. Dugelay, **"LVT Face Dataset: A benchmark dataset for visible and hidden face biometrics"** in *International Conference of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, September 2023.
- **N. Mirabet-Herranz**, J-L. Dugelay, **"Beyond the Visible: Thermal Data for Facial Soft Biometric Estimation"** in *EURASIP Journal on Image and Video Processing*, 2024.

Chapter 5

Chapter 5 introduces a novel structure for facial processing: Transformers. Despite their established state-of-the-art performance in natural language processing and, more recently, in image classification tasks, their application in face biometrics remains unexplored. This Chapter provides an overview of transformers and proposes the first vision transformer trained for facial processing. With it, a single face embedding useful for different facial processing tasks is computed. Building on the positive outcomes from **Chapter 4**, a fusion algorithm between visible and thermal modalities is proposed via the proposed transformer architecture for the estimation of three different soft biometric traits.

Part of the work presented in this chapter was published in:

- **N. Mirabet-Herranz**, C. Galdi, J-L. Dugelay, **"One Embedding to Predict Them All: Visible and Thermal Universal Face Representations for Soft Biometric Estimation via Vision Transformers"** in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, Seattle, USA, June 2024.

Chapter 6

The work presented in **Chapter 6** explores the current trend of digital face beautification via SN filters and its implications for face networks. The widespread adoption of automatic beautification through filters on social media platforms has become increasingly popular. Users employ these face filters to align with societal beauty standards, introducing challenges to the reliability of face recognition and deepfake detectors' final decisions. To better understand this impact, in this Chapter, we propose a method to quantify the impact of each filter by using the Structural Similarity Index to measure their aggressiveness. Additionally, we evaluate the influence of digital beautification on two state-of-the-art face verification networks and three popular deepfake detectors.

Part of the work presented in this chapter was published in:

- **N. Mirabet-Herranz**, C. Galdi, J-L. Dugelay, **"Impact of Digital Face Beautification in Biometrics."** in *EUVIP 2022, 10th European Workshop on Visual Information Processing*, Lisbon, Portugal, September 2022.

- A. Libourel*, S. Husseini*, **N. Mirabet-Herranz***, J-L. Dugelay, **"A Case Study on How Beautification Filters Can Fool Deepfake Detectors"** in *12th International Workshop on Biometrics and Forensics (IWBF)*, Twente, Netherlands, April 2024.

Chapter 7

Following, **Chapter 7** investigates the influence of the beautification filters introduced in **Chapter 6** on various AI-based face analysis technologies, including gender classification, apparent age estimation, weight estimation, and heart rate assessment. Through extensive experiments conducted on our

*All three authors equally contributed to the article

self-beautified datasets, the study concludes that beautification filters have a negative impact on age and weight estimation in terms of overall classification performance. However, their use seems to alleviate some biases introduced during the model's training for gender classifiers. Additionally, the chapter speculates that when applied to facial videos, filters can potentially conceal microsignals, such as those used for heart rate estimation, making social media filters a cost-effective and accessible means for users to safeguard their privacy.

Part of the work presented in this chapter was published in:

- **N. Mirabet-Herranz**, **"Social media filters: Beautification for humans but a critical issue for AI."** in *Science Talks*, January 2024.
- **N. Mirabet-Herranz**, C. Galdi, J-L. Dugelay, **"Facial Biometrics in the Social Media Era: An in-Depth Analysis of the Challenge Posed by Beautification Filters."** in *IEEE Transactions on Biometrics, Behavior, and Identity Science (TBiom)*, 2024.

and won the 3rd price at:

- **"Three Minutes Thesis (3MT) Contest"** organized at *EUSIPCO 2023, 31st European Signal Processing Conference,* Helsinki, Finland, September 2023.

Chapter 2

Facial Image Processing: Advances and Challenges

2.1 Advances in Face Biometrics

Biometrics are biological measurements or physical characteristics that can be used for identifying individuals or providing complementary information about them. The history of Biometrics starts in the 1800s. In that century, criminal offenders who were arrested multiple times would often provide different names to law enforcement authorities. In 1870, Alphonse Bertillon proposed *Anthropometries* (or *Bertillonage*) as a method relying on body measurements, photographs, and physical descriptions for identification. He realized that, even though criminals may change their names, certain aspects of their appearance would remain the same. Despite its global use, its deployment quickly faded as different criminals were found to share similar body measurements. In later years, systems based on different biometric traits were explored for people's identification. In 1892, Sir Francis Galton proposed a fingerprint classification system based on *minutiae*, or *characteristics*, which is still in use today. Following, in 1896, Sir Edward Henry collaborated with Galton to create an efficient method for classifying and storing fingerprint information. In 1936, the ophthalmologist Frank Burch identified differences between human irises and proposed its patterns as a method to recognize individuals. More recently, in the 1980s, the National Institute of Standards and Technology (NIST) created a speech processing group to investigate speaker recognition potentialities.

Face is the most intuitive biometric trait for humans. Unlike iris and fingerprint biometric systems, face recognition does not require costly and high-accuracy acquisition sensors. Furthermore, it does not involve physical

interaction with the end user, facilitating the identification of target subjects from considerable distances without their cooperation. Biometric authentication is a security method that utilizes unique biological traits to verify an individual's identity by comparing physical or behavioral characteristics with stored authentic data in a dataset. Face identification specifically employs facial input data to establish a person's identity, while face authentication uses these traits to confirm that individuals are who they claim to be. If the facial data samples match, authentication is successful. In 1987, Kirby and Sirovich developed an algorithm for face recognition based on principal component analysis [8]. Thanks to this development, a few years later, Turk and Pentland achieved real-time face recognition [9]. Nowadays, face recognition technology is extensively employed for a wide range of applications. In smartphones, it serves as a biometric authentication method and in security scenarios, face recognition is employed for identifying and tracking individuals in high-security areas and for access control systems in buildings.

Humans have demonstrated proficiency at discerning similarities between two face images captured under diverse conditions, but automated face recognition faces several challenges in such scenarios. Variations in age, pose, and illumination as well as changes in appearance due to beautification or facial hair, still pose difficulties. Although AI-based face recognition solutions increasingly improve their accuracy, these gains often come at the cost of highly complex models harder to interpret. Those are also referred as to 'black-box' models. Gaining more insights about such models begins by identifying the influencing factors of AI-based face recognition and comprehending their impact on the overall system performance. Moreover, the integrity of face data has become increasingly compromised in recent years due to the growing popularity of deepfake techniques. Generative Adversarial Networks (GAN)[10] and, more recently, Denoising Diffusion Probabilistic Models (DDPM) [11] have led to significant advances in the generation of synthetic media, namely video deepfakes. These fake videos aim to portray a person in a situation they have not experienced in a highly realistic way to deceive the human eye. Although various types of deepfake videos exist, the currently most prevalent involves a *face swap* between a target person and an individual in a video. While these fake images can be seen as entertainment for the film and advertising industries, they also raise issues of privacy and credibility with the massive sharing on social networks. Moreover, their use in biometric spoofing is escalating, posing a significant threat to the security of these systems.

Table 2.1: Overview of soft biometric modalities and key human features within each category.

Permanent*			***Temporal***	
Global	**Face**	**Body**	**Biological**	**Clothing**
Gender Age Weight Height Microexpressions	Eye color Ethnicity Nose type Lip thickness	Arm length Wrist size Tattoos Gait	Hair style Hair color Facial hair	Head coverage Clothing color Footwear type Eye glasses

*In this context, "*permanent*" refers to a trait unchangeable over a short period, for instance, during at least 24 hours.

2.1.1 Facial Soft Biometrics

Soft biometric traits are human characteristics typically described using human-understandable labels and measurements. Soft biometrics, such as gender, age, height, weight, ethnicity, hair color, etc., are not unique to the individual but can be aggregated to provide discriminative biometric signatures. Indeed, their use has been proposed in the literature to enhance the performance of traditional biometric systems and enable identification based on human descriptions [2]. Although these soft biometric traits have only recently been considered in biometrics, they hold tremendous potential for human identification. Unlike fingerprint or iris, soft biometrics can be obtained at a distance without subject cooperation and from low-quality video footage, making them ideal for use in surveillance applications [12].

In addition, the estimation of various soft biometric traits has been employed as an additional block at the end of an FR pipeline as an explainability technique. The prediction of different traits, namely gender and age, of the probe and gallery face images has been used in cases where face verification fails in order to provide insights into whether there is a correlation between the FR system error and the error of soft biometric estimation networks [C1]. Moreover, soft biometric descriptors are intrinsic to conventional human descriptions [13], as we naturally use these traits to identify and describe each other. Inspired by the classification proposed by previous research [14], Table 2.1 provides an overview of soft biometric modalities and presents examples of associated human traits within each modality. *Permanent* traits, unlike *Temporal* ones, are not modified by a user in a short period. This thesis specifically puts its focus on the estimation of *Permanent* features that can be inferred from face images.

Gender and Age

Regarding the use of the term "gender" in this thesis, since the annotations of the datasets employed were mainly made manually by the researchers who collected the data, we presume that the gender was annotated based on the *assumed gender* of the persons portrayed. When the face images are of celebrities (mainly actors, singers, and athletes), a large majority of the annotations may be in fact the *affirmed gender* of the people portrayed at the time of the data collection. In the following Chapters, we refer the reader to the datasets' source papers for more information about gender annotation.

Antipov *et al.* proposed the first deep learning-based approach for gender estimation [15] from face images. They presented an ensemble model based on three Convolutional Neural Networks (CNNs). After that, multiple works [16], [17] have used deep architectures and local features to design deep neural networks for facial gender recognition. In particular, D' Amelio *et al.* [18] achieved notable success with a model for gender classification from real-world face images where features are extracted through Visual Geometry Group (VGG)-based CNN.

Although age estimation from face images has been an extensively studied topic in the literature [19], age estimation through deep learning models gained traction in 2015 when Wang *et al.* introduced a method employing a CNN architecture consisting of 3 convolutional layers, 2 pooling layers, and 1 fully connected layer to extract features, followed by linear Support Vector Regression (SVR) for age estimation [20]. Inspired by them, other researchers explored various CNN architectures for the same purpose [21], [22]. Rothe *et al.* [23] designed a specialized CNN known as the Deep EXpectation algorithm (DEX) of apparent age. DEX first detects the face in the test image and then extracts the CNN predictions from an ensemble of 20 networks on the cropped face achieving higher performance than previous models. Zhang *et al.* [24] employed a Long Short-Term Memory (LSTM)-based method along with a Residual Network model, constructing an AL-RoR model with 34 layers to extract features for age estimation. In the same line, Wang *et al.* [25] proposed convolutional sparse coding to extract unsupervised learned features of aging, with costumed pooling layers pooling applied to the extracted feature map for better capturing of aging signs.

All those approaches are based on Single Task Learning where only the estimation of one attribute at a time is performed. In a more transversal approach, researchers have explored Multi-Task Learning to jointly learn gender and age at the same time. A successful approach is proposed by Gupta *et al.* who studied different structures for multi-attribute learning with CNN [26].

Some works have initiated the exploration of the thermal spectrum for facial processing models. In the literature, two soft biometrics, namely gender and ethnicity, have been estimated from thermal input data. Chen *et al.*, presented the first work on gender classification from faces in the thermal spectrum [27]. They proposed a pipeline of techniques that consists of Local Binary Pattern method to detect the edges in an image, followed by Principal Component Analysis for dimensionality reduction, and finally, a Support Vector Machine as a classification technique for estimating the subjects' gender. Similarly, Abouelenien *et al.* utilized the Eigenfaces method for visible faces and statistical measurements of pixel color for thermal faces, employing decision trees for gender classification. Fusion between visible and thermal data was integrated within the decision tree model [28]. Deep learning structures began to be explored by Narang *et al.* [29], where they trained a VGG-CNN structure with visible faces and tested it on thermal faces for gender and ethnicity classification. Farooq *et al.* performed transfer learning from nine famous architectures to estimate gender from thermal data [30], including ResNet-50, ResNet-101, Inception-V3, MobileNet-V2, VGG-19, AlexNet, DenseNet-121, DenseNet-20, and EfficientNet-B44. They also proposed GENNet for the same task. More recently, Abdrakhmanova *et al.* proposed a combination of bidirectional recurrent neural network and CNN for extracting features from Visible-Thermal-Audio and then performed fusion at the feature level to estimate the person's gender [31]. They classify females and males afterward with a final decision layer.

In **Chapters 4 and 5** of this thesis, we present the results of two author's publications where a comparison between visible and thermal spectra [J1] and the fusion of both spectra [C2] for soft biometric estimation is performed. The traits considered are gender, age, and weight being a novelty concerning previous work on thermal imagery for biometrics.

Weight

Weight is a soft biometric trait that proves useful for daily health assessment. In contrast to height, weight information has been demonstrated to be discernible in the face, as bone and muscle present in this area exhibit different densities [32]. This thesis places special emphasis on developing an accurate and robust model for weight estimation from face images. Little attention has been given to remote weight estimation from face images in the literature, with existing methods showing several kilograms of error.

Despite the fact that traditional measurement approaches involve physical contact, self-diagnostic image-based methods are becoming a trend nowadays due to the proliferation of high-quality cameras on affordable mobile

phones [33]. Previous work on self-diagnosis has mainly focused on Body Mass Index (BMI) estimation from face images [33]–[35]. However, BMI, being a ratio between weight and squared height, is prone to accumulating errors in its estimation. The height of a subject remains consistent, and with this information, the problem of estimating BMI becomes equivalent to estimating weight.

Weight estimation via a person's image has predominantly focused on full-body images and videos. In 2010, Velardo *et al.* studied the feasibility of predicting the weight of a person accurately from anthropometric data accessible from the subject's image [36]. In their study, they applied multiple regression analysis to the set of anthropometric features obtained from the image. Also utilizing anthropometric data [37], Cao *et al.* presented a copula-based technique aimed at reducing the impact of noise on weight estimation from geometric measurements. In 2012, Labati *et al.* proposed the first, to our knowledge, deep learning approach for image-based weight prediction [38]. They automatically extracted a set of features using image processing from a pair of frame sequences of a walking person, and these features were then processed with a feed-forward neural network and by reaching high performance they demonstrated that artificial intelligence models are effectively able to automatically estimate a person's weight.

However, little attention has been given to direct weight estimation from a subject's face image. Only a handful of research studies have addressed the problem of automated face-based weight estimation. In 2018, Dantcheva *et al.* conducted, for the first time, a study on height, weight, and BMI estimation from a single subject's facial image by implementing a ResNet architecture with 50 layers [39]. Motivated by this work, in 2019, Haritosh *et al.* addressed the challenge by defining a two-step network composed of a feature extraction model plus a customized artificial neural network consisting of 3 fully connected layers. They reported the performance of their model for various feature extractors on two different datasets covering different weight ranges. In 2020, Han *et al.* addressed the impact of the lack of labeled data on a Convolutional Neural Network (CNN) by presenting an auxiliary-task learning framework for weight estimation [32]. To prevent their network from suffering from poor performance due to a lack of labeled data, they defined other features such as age and gender prediction as their auxiliary tasks.

Chapters 3, 4 and 5 of this thesis outlines the advancements made in remote weight estimation from facial images. In various author's publications, networks are proposed for weight estimation from visible [C3] and thermal [C4] face images.

2.1.2 Facial Hidden Biometrics

Faces encode not just information related to a person's identity but also other microsignals that might describe physical and behavioral attributes. Some of those microsignals are seen as health cues therefore their extraction from the biometric data may aid in medical diagnosis. Ross et al. presented a literature review on different health cues embedded in the commonly utilized forms of audio-visual biometric data [3]. According to their classification, there exist three types of health cues that can be extracted from facial shots: 1) Physiological, 2) Psychological and 3) Genetic. Physiological parameters namely BP [40], drug abuse [41] and Heart Rate [42] are the most popular among the ones that can be estimated from faces. Nonetheless, Psychological parameters including apathy [43] or stress [44] and Genetic cues like Noonan Syndrome [45] or Cornelia de Lange [46] have also been studied in the literature.

Among all aspects, physiological parameters such as BP and HR have been a more prominent focus in the literature. The estimation of both parameters is based on Photoplethysmography. Photoplethysmography (PPG) is a low-cost and noninvasive means of sensing the cardiovascular blood volume pulse through subtle color variations in the reflected light of human skin [42], [47]. Although PPG is typically implemented using dedicated light sources, Verkruysse *et al.* [48] showed that using ambient light as an illumination source is sufficient to capture a person's vital signs from RGB videos. Most of the remote PPG (rPPG) extraction algorithms are based on handcrafted features and consist of a two-stage pipeline: first, extracting the rPPG signals from the face, and then performing frequency analysis to estimate the corresponding average HR and/or BP using a peak detection algorithm. Additionally, they require various preprocessing techniques such as skin segmentation, color space transformation, signal decomposition, and filtering steps, among others. Some filters necessitate parameter adjustment and tuning according to the data being used, making these approaches nearly impossible to replicate [49].

Recent works have focused on implementing deep learning techniques for rPPG signal extraction when a large amount of labeled data is available [50]–[54]. Their performance can also be improved by increasing the training set size, unlike previous hand-crafted methods. Some of those researches have focused on extracting the rPPG signal aiming for HR monitoring [50], [52], [55], [56] by using novel techniques such as contrastive learning frameworks [57] and vision transformers [58]. On the other hand, other publications use this signal estimation for BP assessment [40], [59]–[61]. Since the change of facial color intensity caused by cardiovascular activities is weak, environmental illumination changes and subjects' facial movements will produce irregular

noise in rPPG signals, resulting in distortion of heart rate pulse signals and affecting the accuracy of heart rate measurement. The thermal spectrum has been considered for rPPG signal extraction in the context of HR [62] and BP estimation [63].

Deep Learning for Direct HR Estimation

Several works have focused on estimating the HR in beats per minute (bpm) directly from face videos, without an intermediate signal estimation step. Spetlik *et al.* proposed a two-step CNN to directly estimate heart rate from a face video [54]. Wang *et al.* adopted a double feature extraction stream by first applying a low-rank constraint to guide the network to learn a robust feature representation and then extracting rPPG signals [64]. Niu *et al.* introduced a new data transformation to represent both temporal and spatial information in a 2D manner from face videos as input to a deep heart rate estimator [51]. In future research, they refined this approach by using multiple ROI volumes as input [65] and performing data augmentation [68]. Song *et al.* created their own version of spatio-temporal maps constructed from pulse signals extracted from existing rPPG methods to feed their CNN [53]. Hu *et al.* compared the effectiveness of extracting spatial-temporal facial features using 2D-CNN against 3D-CNN [70], while Lokendra *et al.* experimented with the utilization of Action Units (AUs) and Temporal Convolutional Networks (TCN) for denoising temporal signals and improving HR estimation [69]. Huang *et al.* proposed a deep neural network consisting of 2D convolutional layers and LSTM operations [67]. Bousefsaf *et al.* presented a method relying on 3D networks with embedded synthetic signals in real videos [66], which outputs values recorded in a histogram composed of intervals of 2.5 bpm. However, this network has been noted for its slow processing time. More recently, Li *et al.* designed a three-dimensional spatio-temporal stack convolution module with a multi-hierarchical feature fusion module to strengthen the spatio-temporal correlation of multi-channel features [71].

Despite the progress in extracting rPPG signals and estimating HR from facial videos using deep learning, little attention has been given to establishing unified criteria for choosing input data, especially for learning-based approaches. We aim to address this gap by providing a study on the choices that authors have to make when implementing a deep learning HR estimator. To this end, we provide an overview of recent works using deep learning structures for direct HR estimation from face videos in Table 2.2. The table presents the model structure chosen in each approach, the type of input data passed to the network, the ROI selected, and the public dataset (if any) in which their results were reported. As shown in Table 2.2, no comparison

Table 2.2: Overview of relevant works aiming for a direct HR estimation from face videos via deep learning structures. Includes model structure, type of input, ROI selected, test dataset and metrics reported.

Paper	Year	Structure	Input data	ROI	Datasets	Metrics
HR-CNN [54]	2018	CNN	RGB	Full frame	COHFACE MAHNOB PURE ECG-Fitness	MAE RMSE ρ
SynRhythm [51]	2018	ResNet18	Spatial-temporal maps	Nose and cheeks	MAHNOB MMSE-HR	Me STDe RMSE MER
2-stream CNN [64]	2019	Two layer LSTM	Spatial-temporal maps	Full frame	COHFACE PURE	MAE RMSE ρ
RhythmNet [65]	2019	ResNet18	Spatial-temporal maps	Full face	MAHNOB MMSE-HR	Me STDe MAE RSME MER ρ
3D-Mapping [66]	2019	3DCNN	Shuffled G pixels	Full frame	UBFC-RPPG	Me STDe MAE RMSE
Visual-CNN [67]	2020	CONV2D + LSTM	RGB	Full face	-	STDe MAE RSME MER ρ
Robust-CNN [68]	2020	CNN	Spatial-temporal maps	Full face	MMSE-HR	STDe MAE RSME MER ρ
HR-CNN [53]	2020	ResNet18	Spatial-temporal maps	Nose and cheeks	MAHNOB ECG-FITNESS	Me STDe MAE RSME MER ρ
AND-rPPG [69]	2021	Temporal CNN	RGB	Full face	COHFACE UBFC-rPPG	STDe MAE RSME ρ
rPPGNet [70]	2021	**2D vs 3D CNN**	RGB	Full frame	COHFACE PURE	Me STDe MAE RSME
3D-CNN [C5]	2022	3D-CNN	**RGB vs R vs G vs B**	**Full face vs cheeks vs forehead**	COHFACE	Me STDe MAE RSME MER ρ
3D-Attention [71]	2023	3D-CNN	RGB	Full face	COHFACE	STDe MAE RSME ρ

between input data or ROI is done when the selection of those needs to be made. The only comparative study made, up to our knowledge, concerns the performance of HR when a 2D or a 3D CNN is selected as the network [70].

In **Chapter 3**, a 3D-CNN for HR estimation is proposed [C5] which was also the result of an author's publication. In addition, we aim to enlarge the state of the art by covering studies such as a comparison of different facial areas to be selected as ROI (full face, cheeks, and forehead) and which channels of the input video provide the most valuable information for a CNN-based HR estimator.

2.2 Challenges in the Digital Era

2.2.1 Social Media Filters

The number of online users has increased by over a billion in the last few years. More than 18.2 million text messages are transmitted in a minute and a big portion of them are multimedia content [72]. In addition, the act of uploading facial pictures to the internet has become more common since the use of technology and social media platforms in particular, also known as Social Networks (SNs), has grown. For instance, users take and share their face images on SN aiming to attract the public through visually appealing and engaging content [73]. *Selfies*, photographs that a subject takes of oneself, have been studied in the literature as a positive way to allow people to express themselves [74]. Nevertheless, posting selfies on SNs can also bring some negative effects on users. They may, for instance, feel dissatisfied with their appearance, especially when compared with other images that are often artificially enhanced to conform to current beauty standards.

Since its creation in 2004, Facebook has been the most popular social media platform, nevertheless, from 2017 onwards, surveys indicate that young users have shifted towards YouTube, Instagram, Snapchat, and TikTok [75]. These popular SNs, offer to their users filters that automatically apply modifications to their face images such as the alteration of different facial traits to beautify their facial appearance. About 600 million people use filters monthly on Instagram or Facebook, and 76% of Snapchat users use them every day [76]. Filtered images have been proven to be among the most heavily engaged photos on SNs [77], augmenting the popularity of these techniques. Filters are not necessarily used to compromise face analysis models but they can potentially disrupt their capability of classifying identities or estimating health traits.

The different effects of filters can range from simple color image transfor-

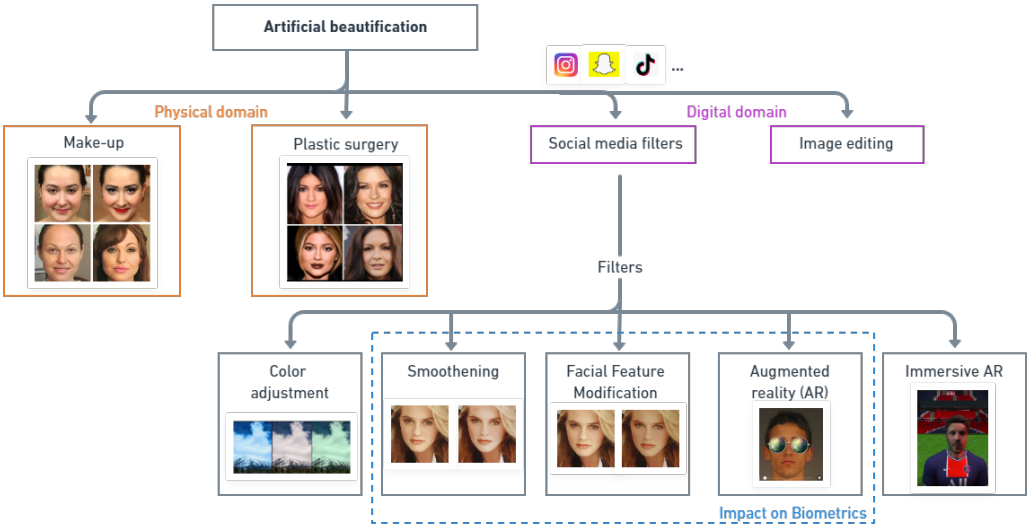


Figure 2.1: Beautification techniques are classified based on their application in either the physical or digital domain.

mations to the addition of virtual elements to the scene. Applying filters to users' content has become a widespread practice since filtered images have the largest engagement in social media [77]. Moreover, the application of filters to facial multimedia does not require any prior expertise, in opposition to other image editing techniques, making them highly accessible to the average social media user.

As represented in Figure 2.1, in this thesis a classification of the filters available in social media is proposed:

- Colour Adjustment Filters (CAF): They include color changes to the captured image/video. They apply to the color and luminance channels or consist of converting the image/video from RGB to black-and-white modality.

- Smoothing Filters (SF): Smooth and blurry face skin, as if a foundation layer was applied. They focus on skin pixels, leaving the rest of the multimedia regions untouched.

- Facial Features Modification Filters (FFMF): They modify the biometric features of the user's face. Some of the most common effects are the enlarging, shrinking, and sharpening of different facial lines (e.g.: augmenting the eyes or the lips size, defining the nose lines and modifying the eyebrows position). The use of FFMF is not necessarily noticeable by the human eye since the alteration of facial features is done unnoticed.

This type of filter adds no extraneous elements or textures to the user's face.

- Augmented Reality Filters (ARF): An Augmented Reality (AR) filter is a mask-like filter that adds unexisting elements to the scene in real time. They allow the user to incorporate different gadgets on their heads/faces such as animal ears, superhero masks, and flowery elements but also to see how a specific product might look on their faces (e.g. eyeglasses).
- Immersive AR Face Filter (IARFF): They place, in real-time, the users' face or some elements of it (e.g.: eyes, mouth) into a virtual digital object/scene.

FFMFs are often applied to the user's face for beautification. As mentioned above, they subtly modify a subject's face, those differences being difficult to spot to the naked eye without a reference picture. These filters are commonly applied in conjunction with SF and a specific type of ARF that adds a layer of makeup to a person's face. Other social media filters affect the whole image and their modifications are more visible, making it easier to differentiate an image that has been beautified from an unchanged one. CAF is frequently applied in landscape pictures while ARF and IARFF play an amusement role.

The filters studied in Part II are selected among those available on the most popular social media platforms and from those that can be applied to existing multimedia content. Primarily, FFMFs are studied as the changes they apply are usually unnoticeable by the human eye if no direct comparison with the original image is made and therefore an image of this kind can pass as unaltered during a visual examination.

2.2.2 Face Beautification

Beautification is the visual alteration of the perceived shape and texture of a human face to enhance the subject's appearance. In the past decades, experiments showed that beauty can be assessed and modified since an objective component in beautification appears to be linked to the proportions of features [78]. Thus, researchers have sought to predict the way humans perceive facial aesthetics [79]. Facial beautification comprises the manipulation of biometric traits which has the potential to impact the features extracted by automatic facial processing systems, for instance compromising the effectiveness of FR systems in security applications and the utility of e-health models, including remote heart rate HR estimation. Beautification techniques can be implemented in the physical domain through practices

like makeup or plastic surgery, or in the digital domain using tools such as image editing software and social media filters. Figure 2.1 illustrates this categorization.

Makeup

Facial cosmetics can enhance or disguise facial traits since the use of makeup can visually modify the proportions of different facial characteristics, such as the eyes and cheekbones. In 2010, Ueda *et al.* observed that light makeup makes it easier to recognize a face, while heavy makeup makes it more challenging [80]. Moreover, makeup has been proven to be an effective attack for FR systems [81]. In this type of attack, the attacker might apply a substantially high amount of cosmetics to emulate the facial appearance of a target user. Chen *et al.* explored the impact that such modifications have on other traits, namely gender and age [82]. In recent studies, the vulnerability of different open-source FR systems, such as ArcFace [83], is assessed concerning various makeup presentation attacks. Eckert *et al.* created a dataset containing multiple images per person with and without applied cosmetics and conducted preliminary tests to evaluate their impact on automatic face recognition [84].

Other works have evaluated integrating makeup detection schemes into biometric systems to improve FR models [85] by proposing a dynamic weighting of the extracted traits for FR according to the makeup classification result. As a next step, Rathgeb *et al.* present in their work a makeup attack detection scheme based on comparisons between face depth data and face depth reconstructions obtained from RGB images of potential makeup presentation attacks [81].

Plastic surgery

Facial plastic surgery is commonly used to correct feature imperfections or improve the visual appearance of a subject by removing birthmarks and scars, enhancing desired traits, and correcting asymmetric features. Plastic surgery objectives can be divided into two classes: *reconstructive* plastic surgery, which aims to rectify various facial anomalies, and *cosmetic* plastic surgery, which enhances the visual appearance of facial structures.

When a subject undergoes plastic surgery, both the shape and texture of facial features are changed to varying degrees, altering their appearance. As a result, existing identities may become unknown to the already existing FR system and their reference templates. These surgical changes pose a challenge to FR technology [86], [87]. Facial plastic surgery is usually employed benignly

to improve a person's appearance, but research has pointed out the use of plastic surgery by criminals to "manipulate" their facial identity with the intent to deceive FR systems [87]–[89]. This increases the challenges faced by FR technology, which not only needs to preserve classification accuracy but also be robust to the changes produced by this beautification technique.

Image editing

Digital face retouching is a commonly used application accessible on a broad range of devices, including mobile phones, tablets, and personal computers. Rathgeb *et al.* presented in their survey [86] a first categorization of digital beautification types and new challenges created by beauty face distortions. In addition, they give attention to facial retouching in the *digital* domain with different software. Bharati *et al.* reported high errors in FR SotA systems when faces were digitally modified [90]. Kose *et al.* [91], simulated nose alterations traditionally achieved with the use of makeup and/or plastic surgery. Their results reported as well a loss of precision in FR models for both 2D and 3D faces. In the past years, some works have focused on automatically beautifying faces without interfering with the identity of the subject. Novel methods for digital face beautification have been presented to increase the predicted attractiveness of a subject, maintaining a strong similarity between the pre- and post-modified face [92], [93]. In more recent studies, Diamant *et al.* trained a GAN conditioned on a beauty score [94] while Zhou *et al.* adapted the conditional GAN by adding the identity feature as another condition [95].

Social media filters

Among image editing applications, this thesis focuses on analyzing automatic filtering via social media tools. Due to the trend of social media face filters, on the one hand, some researchers are directing their studies toward creating facial filters, such as skin smoothing, that are undetectable to the human eye [96]. On the other hand, some works have started addressing the impact of filters on facial processing tasks. Digital filtering has an impact on identity recognition, as freckles or other person-specific facial marks can disappear. This becomes particularly problematic when crucial face regions are occluded, for example, with the use of ARF. Hedman *et al.* assessed the impact of CAF and ARF on FR and proposed a counter-filter based on a modified version of the U-NET segmentation network [97]. To improve the FR system, deep learning algorithms and distance measures are applied to the features extracted using a ResNet-34 network. Botezatu *et al.* studied

Table 2.3: Works addressing the impact of beautification in different manners. We select a representative set of papers based on the date of publication and number of citations. We present three pieces of research per beautification type.

Beaut. Type	**Work**	**Analysis performed**					**Conclusions**
		Face	**Gender**	**Age**	**Weight**	**HR**	
Makeup	[80]	x	x				Light makeup increases FR rates
	[82]		x	x			Gender spoofing and age alterations via makeup are possible
	[81]	x					Simple makeup is of no risk for up-to-date FR systems
P. Surgery	[88]	x					Strong negative impact on FR
	[89]	x					Tradeoff between plastic surgery users and impostors for FR
	[87]	x					Surgery presents a problem for FR algorithms that are texture-based
Digital	[86]	x					Digitally facial traits can be more substantially altered
	[90]	x					Digital retouching leads to high errors in FR
	[93]	x					GAN beautification can ensure identity after beautification
Filters	[99]	x					Small impact of filters on FR
	[97]	x					Big impact of filters if the eye region is occluded by AR items
	[C6]	x	x		x		Filters usually penalize but can increase model performance
	[J2]	x	x	x	x	x	Moderate impact on FR and Age Improve Gender performance Heavily penalize Weight and HR
	[C7]	x					Deepfake detectors suffer from a lack of performance

the impact of ARF, also called "fun" selfie filters [98]. The authors evaluated their impact on different face detector and recognition models and proposed a GAN-based filter removal algorithm. Although ARF adds artificial elements that act as occlusion, thus removing essential information from the face image, the risk to biometric systems is minor since a face of this kind is easily detected as processed through visual inspection, for example, in a border control scenario. In a more socially-oriented approach, Riccio *et al.* drew key insights such as the discovery of a general homogenization of the beautified faces when compared to the original ones [99]. They argued that the application of beauty filters does not have a significant impact on state-of-the-art FR models, as filters attempt to preserve the user's identity.

In Table 2.3, a compendium of relevant works assessing the impact of all types of beautification is presented. Brief conclusions extracted from these articles are provided as well as the facial processing technique(s) analyzed. The author's contributions are positioned in the Table and will be presented in detail in **Chapters 6** and **7**.

In Part II, a comprehensive study of the impact of different social media beautification filters on an extensive number of up-to-date facial processing techniques is presented, which in addition to the above-mentioned publications, is the result of a third prize in the 3MT contest organized yearly at EUSIPCO [P1] and a a video journal publication [V1].

Part I

Soft Biometric Estimation

Chapter 3

Facial Soft Biometric Estimation in the Visible Domain

This chapter aims to present improved AI-based models for facial processing, beyond established techniques such as FR and traditional soft biometrics extraction. Identification of individuals, as well as the estimation of gender and age, has been a focal point in research for decades, achieving these models high accuracy even in challenging conditions. This is why, in this chapter, the aim is placed on achieving high-accuracy models for the remote estimation of weight and heart rate from a person's face.

Section 3.1 motivates the estimation of weight and HR from face visuals. Subsequently, Section 3.2 presents the proposed pipeline for weight estimation from face images and its results, while Section 3.3 introduces the HR estimation model and presents its performance. Finally, Section 3.4 summarizes the contributions of the chapter.

The research question that this chapter aims to answer is: **RQ1**, i.e. *What are the most suitable deep learning architectures and training methodologies for various facial processing tasks?*

3.1 Introduction

As a branch of biometrics research, the origin of soft biometrics can be traced back to the need for non-intrusive solutions to extract the physiological traits of a person. Traditionally, the objective of using facial soft biometrics is to facilitate face recognition systems rather than developing standalone facial soft biometrics recognition systems [14]. In recent years, a large number

of research experiments have been conducted to estimate gender, age, and ethnicity from images or videos captured in constrained or unconstrained environments. Furthermore, the combination of these three global traits has been used for identification. However, nowadays, the estimation of biometric features is interesting beyond people's identification.

Among all soft biometric traits, weight is also an indicator of both physical aspects and health conditions. Unlike gender and height, body weight changes during a person's adult life and needs to be periodically measured. Conventional weight measurement techniques require the cooperation of the subject to be measured, which might not be possible during medical emergencies, road accidents, or due to different patient disabilities. In non-cooperative scenarios, visual estimation of the patient's weight by a health professional is preferred [100], but these estimations might not always be accurate [36].

In addition, accurate remote pulse rate measurement from RGB face videos has gained significant attention in recent years. This technology enables non-invasive and contactless monitoring of a subject's heart rate, proving useful in numerous potential applications. There is currently a global trend towards monitoring e-health parameters without the use of physical devices, facilitating at-home daily monitoring and telehealth. rPPG technologies allow for non-intrusive measurements, which are highly relevant when contact must be prevented (e.g., to avoid skin damage) or when users' cooperation cannot be required (e.g., in surveillance scenarios). Several studies have demonstrated that a laptop camera is sufficient to capture the subtle changes in skin color necessary for successful heart rate estimation [47], making this technology accessible to individuals with a webcam-equipped laptop or a mobile phone.

3.2 Weight Estimation from Face Images

In this section, we define an optimal transfer learning protocol for a Residual Neural Network with 50 layers (ResNet50) architecture, achieving better performance than the current state-of-the-art proposals. Additionally, we demonstrate the crucial roles of gender-splitting, image cropping, and hair occlusion in weight estimation, factors that may not necessarily play the same role in face recognition. We use up-to-date explainability tools to illustrate and validate our assumptions. To ensure a fair comparison with other approaches, we conduct extensive simulations on the most popular publicly available face dataset annotated by weight. Moreover, we aim to address the limitations of this dataset by introducing our self-collected dataset, consisting of 400 new face images.

Table 3.1: Evaluation of the proposed ResNet50 for weight estimation in the VIP_attribute dataset with and without Image Augmentation.

	MAE	ρ
No Image Augmentation	9.23	0.63
Image Augmentation	**7.54**	**0.78**

3.2.1 Methodology

It is well-known that transfer learning contributes significantly to enhancing the performance of AI-based algorithms for a variety of tasks in computer vision. However, achieving optimal results requires the application of a suitable transfer learning strategy during the training process. In the following paragraphs, we detail the various techniques that constitute our weight estimation pipeline, namely image augmentation, transfer learning, and the prioritization of gender-based networks.

Image augmentation

Image data augmentation techniques play a crucial role in creating CNN that are invariant to object location, distortion, and image brightness, thereby enhancing the network's ability to generalize. In this work, we applied augmentation techniques to the training dataset to enrich it with new, plausible examples. These variations of the training set images aim to be representative of other testing samples that the model may encounter. For augmentation, we utilized the Python library Augmentor [101], which randomly flips and distorts input face images and adjusts their color, contrast, and brightness. The inclusion of these augmentation techniques resulted in a decrease in the test Mean Absolute Error (MAE) from 9.23 to 7.54 kilograms (kg) and an increase of the Pearson's correlation coefficients (ρ) between the ground truth and the predicted values, as displayed in Table 3.1.

Transfer learning

Transfer learning techniques have proven effective in various real-world applications, often surpassing the results obtained by training an entire CNN with random initialization [102]. In our work, we aim to validate whether applying transfer learning from an analogous task, specifically age estimation from face images, yields significant benefits for our task. This serves as a crucial step towards achieving optimal remote weight estimation.

In a CNN, filters operating directly on the input data learn, during the training process, how to extract low-level features such as edges. As layers are

Table 3.2: Evaluation of the proposed ResNet50 for weight estimation in the VIP_attribute dataset for a different number of frozen hidden layers (hl) and cost functions.

	MAE		**MSE**		**HUBER**	
# hl	MAE	ρ	MAE	ρ	MAE	ρ
5	9.25	0.67	9.97	0.61	9.37	0.69
10	8.5	0.72	9	0.66	9.62	0.71
15	8.57	0.71	8.63	0.66	8.17	0.72
20	8.23	0.75	7.79	0.75	**7.54**	**0.78**
25	9.15	0.68	8.21	0.69	8.8	0.72
30	9.63	0.68	9.38	0.70	9.28	0.72

stacked, the network can abstract more complex traits, leveraging increased model depth. In our approach, we retain the initial layers of a pre-trained ResNet50, as they extract general facial shapes, and thus, we do not retrain them. Instead, we focus on adapting the last set of layers, originally designed for age estimation, to our current weight estimation objective. We conducted a comprehensive study to determine the most suitable number of layers to freeze during transfer learning. Additionally, we explored different types of cost functions for the training process, which results are detailed in Table 3.2. Our findings indicate that freezing the first 20 layers of the ResNet50 architecture, trained with the Huber loss function, yields the best network performance. This choice aligns with the desirable properties of the Huber cost function. The presence of outliers in the VIP_dataset poses a challenge to the Mean Squared Error (MSE) cost function, as large errors heavily impact model training. Conversely, the MAE cost function proves sensitive to local minima.

Gender-based network

Due to differences in bone mineral and muscle density between male and female bodies, their facial appearances can vary even when they share the same weight [32]. Our objective is to illustrate that weight estimation can benefit from gender perception. To demonstrate this, we implement both gender-mixed and two gender-based ResNet50 models. This experimentation aims to show that, unlike many state-of-the-art face recognition models, a weight estimator is significantly influenced by prior gender classification. However, it is important to note that various factors such as muscular density, percentage of water, or the use of medication can also impact facial appearance and, consequently, weight estimation. We have chosen gender as our primary

factor of study, recognizing that specific medically annotated datasets would be required to further study those additional characteristics.

3.2.2 Experimental Setup

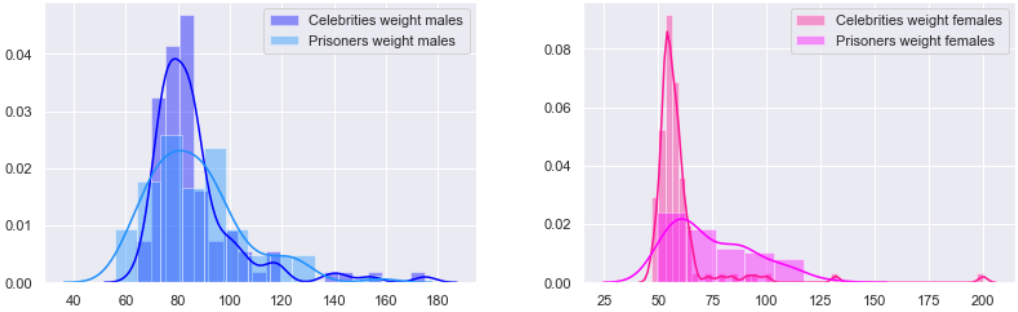
Datasets

Experiments are conducted on two datasets: The VIP_attribute dataset [39], which is the largest publicly available face dataset annotated by weight and our self-collected Prisoners dataset [D1].

VIP_attribute dataset: This dataset comprises 513 female and 513 male face images of various celebrities, including actors, singers, and athletes, collected from the Web. The images consist of frontal views captured under different illumination, expressions, image quality, and resolution conditions. The VIP_attribute dataset is distributed in an already cropped version since the authors applied the Viola-Jones algorithm [103] for face detection to all face samples. In their work, the authors of the dataset noted that weight annotations may not be fully accurate due to inaccurate self-reports or weight fluctuations over time [39]. Additionally, the dataset presents challenges such as artificial beautification techniques like makeup or plastic surgery. To create a balanced evaluation, we performed a random gender-balanced split, resulting in training subsets with 820 individuals and testing subsets with 206 individuals.

Prisoners dataset [D1]: The motivation behind creating a new dataset of face images with associated weight annotations comes from the limited availability of publicly accessible datasets and the restricted range of weight variations covered by them. The VIP_attribute dataset stands out as the largest face dataset annotated with weight. However, it is important to note that celebrities featured in such datasets may not be representative of the general population. To address this limitation and contribute to the existing literature with a more diverse test group reflecting a broader weight distribution, we introduce the Prisoners dataset. Comprising 400 face images (304 male and 98 female), this dataset includes annotations for age, height, weight, ethnicity, gender, eye color, and hair color. The data was collected from the Polk County Jail official webpage[1]. Figure 3.1 presents a comparison of weight distributions between the VIP_attribute and Prisoners datasets for both female and male subjects. Unlike the VIP_attribute, the weight annotations for each prisoner in our dataset are highly reliable. The weights were recorded at the same time the facial pictures were taken, ensuring the accuracy of the annotations and reducing uncertainty in our final predictions.

[1]https://www.polksheriff.org/detention/jail-inquiry



Male subjects' weight distribution. Female subjects' weight distribution.

Figure 3.1: Comparison between the weight distributions of the VIP_attribute and Prisoners datasets.

To evaluate the model, we performed a random split of the subjects into training (320 individuals) and testing (80 individuals) subsets.

Metrics

In Section 3.2.3 the weight estimation performance is reported through the following metrics: The MAE in kg, the Pearson's correlation coefficients (ρ) and the Percentage of Acceptable Predictions (PAP). The PAP was introduced by [36] and it represents the percentage of the prediction whose error is smaller than 10% of the initial weight, i.e. a reasonable error in medical applications.

Implementation details

We resize the face images to 256×256 and provide them as input for our CNN. To boost the learning process, we apply augmentation techniques to each training image for every training epoch. The ResNet50 structures were implemented in TensorFlow and Keras. We initiated the weights of the network with the filters generated by an age classifier [23], trained in more than 20000 images from the UTKFace dataset [104]. The first 20 layers of the model were frozen, the trainable layers were trained during 10 epochs and the final regression layer during 10 epochs more. Adam optimizer was selected with a learning rate set to 0.01. The loss function selected was Huber loss with $\delta = 1$.

3.2.3 Results

We conducted three experiments:

1. Training a single network with the complete training set;
2. Training two separate networks by dividing our training and testing sets based on gender;
3. Conducting cross-dataset and intra-dataset experiments using the Prisoners dataset.

In Table 3.3, we present a comparison between our proposed approach and other weight estimators. The results demonstrate a significant reduction in MAE on the test set for both gender-mixed and gender-based models, achieving a relative improvement of 11.4% and 15.3%, respectively, compared to [39]. Notably, our model achieves an error of 4.24 kg for female subjects, nearly half of the one reported by [39] (8.06 kg and 8.25 kg for female and male subjects, respectively). The VIP_attribute weight distribution has a mean (μ) of 72.60 and a standard deviation (σ) of 21.94. The weights of female and male subjects follow distributions with means $\mu = 58.34$ and $\mu = 86.93$, and standard deviations $\sigma = 11.02$ and $\sigma = 21.00$, respectively. Our results, 4.24 kg (female) and 9.59 kg (male), are more consistent with the weight distribution. Our gender-based model achieves the lowest MAE of 6.91 kg on the VIP_attribute, indicating that training two separate networks helps the model learn appropriate features for weight prediction.

In Table 3.4, the results of experiments conducted on the Prisoners dataset are displayed. When considering the network trained and tested on the VIP_attribute as a baseline, it is apparent that the ResNet50 trained on the VIP_attribute can generalize its predictions to an unknown population, such as prisoners. When no prisoner data is included in the training process, the test MAE experiences a slight increase of 0.41 kg. However, the model trained on the Prisoners dataset struggles to extrapolate its knowledge to the VIP_attribute population, resulting in a test MAE increase of 4.78 kg. It is worth noting that all prisoner pictures were taken in a controlled environment with consistent lighting and background conditions, while celebrity images in the VIP_attribute dataset were collected from diverse sources. Deep learning models tend to extrapolate knowledge more effectively when trained on a variety of images and larger datasets [105] leading to more accurate predictions when training the proposed ResNet50 in the VIP_attribute dataset.

3.2.4 Explainability Study of Clue Factors

Nowadays machine learning models are capable of achieving high predictive accuracy becoming a widespread tool for several applications such as image classification. Although in many cases their performance can be compared

Table 3.3: Performance comparison between the proposed ResNet50 training protocol and SotA networks for face-based weight estimation on the VIP_attribute dataset.

	MAE_f	MAE_m	MAE_{all}	ρ
Dantcheva *et al.* [39]	**-**	-	8,51	0,75
Dantcheva *et al.* [39]	8,06	**8,25**	8,15	0,77
Han *et al.* [32]	-	-	7,20	-
Gender-mixed [C3]	-	-	7,54	**0,78**
Gender-based [C3]	**4,24**	9,59	**6,91**	**0,78**

Table 3.4: Results of the proposed ResNet50 protocol for the intra-dataset and cross-dataset experiments on the VIP_attribute and Prisoners datasets.

	MAE	ρ	PAP (%)
Train: Prisoners, Test: Prisoners	9,79	0,44	46,25 %
Train: VIP, Test: Prisoners	10.20	0,42	43,75 %
Train: Prisoners, Test: VIP	12,32	0,41	36,58 %
Train: VIP, Test: VIP	**7,54**	**0,78**	**62,43**%

with human abilities, it is often at the cost of limited explainability. Simple networks benefit from higher explainability than complex ones. Building transparent deep learning models is the final goal to reach since black-box models do not infer trust for different applications such as healthcare. A general user might be still skeptical when facing black-box approaches in cases where model explainability is a concern. In this subsection, our goal is to enhance our understanding of the decision-making process within the ResNet50 structure by applying explainability techniques. We aim to gain insights that will help us enhance the performance of the weight estimation system.

Different approaches exist for better balancing this trade-off between complexity and interoperability. Techniques such as filter visualization [106], disentangling CNN representations into other structures such as decision trees [107] or directly addressing the learning of disentangled representations where the middle layers are no longer a black box [108] have proved successful. Recently, a new research direction has been proposed aiming at quantifying the contribution of each input feature to the decision taken by the model. Following this research line, we will apply to our ResNet50 two model agnostic techniques: LIME and SHAP.

- In 2016, Ribiero *et al.* proposed LIME (Local Interpretable Model-

Agnostic Explanations), a technique that could explain any classifier predictions by learning an interpretable model around the prediction of a single data instance via input data perturbations [109]. More specifically, LIME modifies a test data instance by altering its input values, in our case image pixels, and observes the impact on the model-outputted weight prediction.

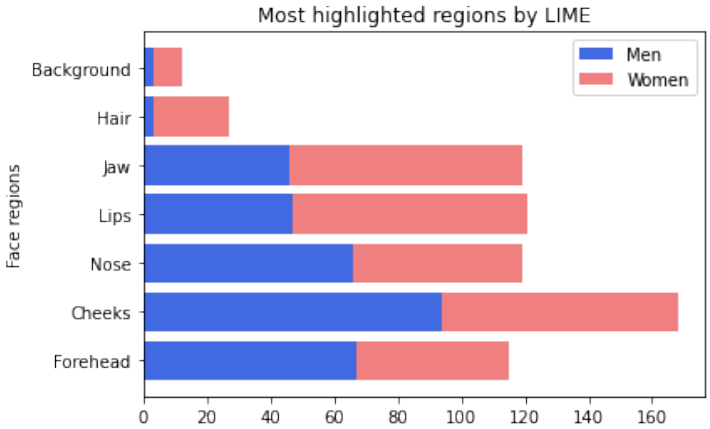
- In 2017, Lundberg and Lee presented SHapley Additive exPlanations (SHAP) [110], a framework based on Shapley values, which refers in game theory to the average of all the marginal contributions to all possible coalitions. In our case, our game reproduces the weight outcome of the model and our players refer to the image pixels.

Visualization of the Explanations

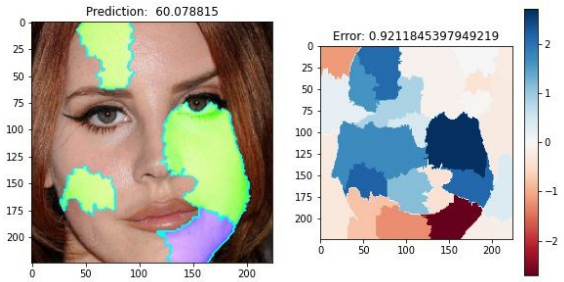
In Figure 3.2 we present representations of the output of LIME (b) and SHAP (c) when those explainability techniques are applied to the face images. In (b), the highlighted green areas represent the image regions contributing to an increase of the weight and opposite to them, the purple ones constitute the portions decreasing the weight value. In (c), the red dots represent meaningful players for the game outcome. The explainability techniques do not assess the validity of the result, instead, they give complementary information on which image areas were most significant for the prediction. Nearly all the highlighted pixels for both methods lay in the face skin areas of the image, excluding regions such as background or eye pupil, reinforcing our model trust since meaningful parts of the image were taken into account. We made a count of the most returned face areas by LIME across the test set and presented them in Figure 3.2 (a) for the male and female subjects. We observed how in all cases, the cheek area was the most highlighted feature. We also noticed how different areas were highlighted for males and females, as is the case of the jaw for women confirming that the model focuses on different face regions depending on gender. Finally, we also counted the times that a non-facial region was highlighted. The background was not often considered relevant information while the hair areas were wrongly taken into account in more cases.

Study Derived from the Explainability Maps

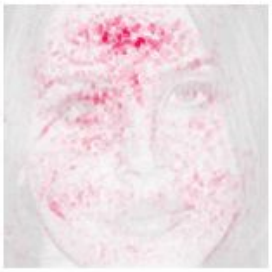
Two additional studies are performed, taking into account the output of LIME:



(a) Count of the most contributive regions after applying LIME.



(b) LIME example



(c) SHAP example

Figure 3.2: Model agnostic LIME and SHAP explainability approaches applied to the proposed ResNet50 in the VIP_attribute dataset.

1. Study on the optimal face cropping as a pre-processing step for the proposed ResNet50 model for weight estimation;

2. Assessment on the impact of hair occlusions on weight estimation from face images.

1. Face detection margins. While assessing trust in the predictions through LIME, we discover that significant regions for our weight estimation model such as the face contour (forehead, cheeks and jaw) are usually excluded from face cropping algorithms since eyes contain the most meaningful information for face recognition tasks [111]. Indeed, the VIP_attribute dataset [39] is distributed in an already cropped version with a narrow bounding box, often excluding parts of the face contour such as the forehead, cheeks, and jaw. To address this limitation, we contacted the authors of the VIP_attribute dataset, who provided us with the original version. Subsequently, we evaluated whether

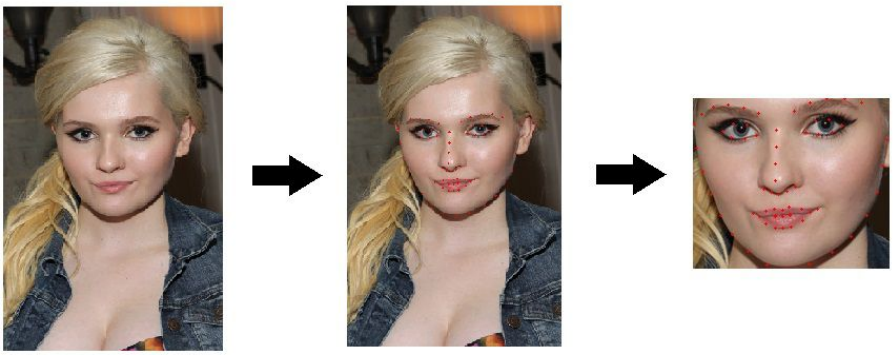


Figure 3.3: Example of the customized face cropping from facial landmarks for the first subject in the VIP_attribute.

employing different croppings, especially those considering larger face areas, would result in a more accurate prediction, as suggested by explainability approaches. In our experiment, we have trained and tested our network for 4 different face cropping methods: Viola-Jones [103], Multi-Cascade CNN (MTCNN) [112], the dlib python package and a customized cropping. We defined our face cropping by considering the highest, lowest, and furthest at the left and furthest at the right facial landmarks computed by the dlib landmark detector. An example of the output cropped images is presented in Figure 3.3. The results in Figure 3.4 illustrate the MAE on the y-axis for various cropping margins on the x-axis. A margin of -0.1 indicates that 10 % of the image height has been subtracted from the top, the bottom, the right and the left of it while a margin of 0.1 indicates that 10 % more of the image has been considered in every direction. Notably, an increased margin of 0.1 shows a clear benefit, particularly for rectangular face croppings (MTCNN and customized), which better adapt to the face shape. However, it is important to note that larger croppings include more hair and background regions, leading to an increased MAE for the network. In the case of our customized cropping, negative margins highly increase the error due to the fact that the upper landmark is located in the eyebrows position, leading to a removal of meaningful information such as eyes when part of the image is subtracted.

2. Hairstyle. Another significant occlusion factor when considering a face image is hair. LIME highlighted hair areas as relevant in some cases, especially those present above the forehead and cheeks. To further investigate the impact of hair on weight estimation, we expanded the annotation of the VIP_attribute dataset. For each subject, we added annotations regarding

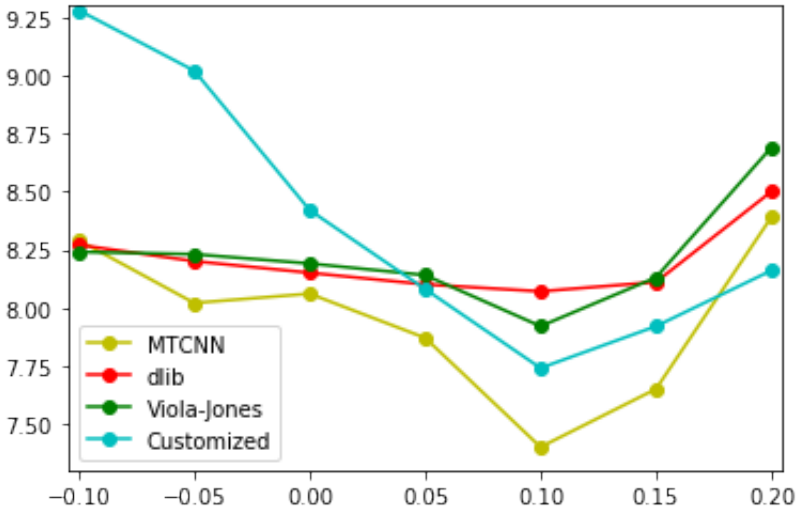


Figure 3.4: MAE in kg of our ResNet50 model for weight estimation in the VIP_attribute test set for various face detectors and cropping margins.

their hairstyle, the presence and type of facial hair, and the presence of glasses. This expansion enables more in-depth studies of these categories. The new metadata is based on the annotation proposed by [113] and is available upon request.

Table 3.5 presents the MAE and MAPE per category. Some categories, such as "Bold-Short Bold," are underrepresented, thus the presence of outliers may influence the high values. However, in the case of facial hair, we observe that the presence of a beard significantly impacts the prediction, increasing the error by more than 2 kg on average, whereas the presence of fringe as an occlusion is not as meaningful for the prediction.

3.3 Heart Rate Estimation from Face Videos

In this Section, we extensively tested a new framework to better understand several open questions in facial HR estimation: which areas of the face are the most relevant, how to manage video color components and which performances are possible to reach on a publicly relevant dataset. From this study, we extract key elements to design an optimal, up-to-date and reproducible framework that can be used as a baseline for accurately estimating the heart rate of a human subject, in particular from the cheek area using the green (G) channel of a RGB video. The results obtained in the public dataset COHFACE support our input data choices and our 3D-CNN structure as optimal for a

Table 3.5: MAE and MAPE of the proposed ResNet50 for weight estimation in the VIP_attribute test set for different hairstyles

	Type	**# of subjects**	**MAE**	**MAPE**
Hairstyle	Bold - Short bold	11	11.32	12.98
	Short	96	8.73	11.12
	Medium	26	5.47	8.00
	Long - Long volume	72	6.12	8.92
Facial Hair	Clean	136	7.40	10.17
	Moustache - Goatee	11	7.47	8.84
	Beard	47	8.30	10.16
	Fringe	11	6.10	9.28

remote HR estimation.

3.3.1 Methodology

"The following paragraphs describe the various methods defined in this thesis for an optimal remote heart rate estimation pipeline from face videos. In particular, they discuss techniques for the selection and comparison of different ROIs, the rationale behind using single-channel vs. multi-channel inputs for a CNN architecture, and the potential of 3D CNNs for the studied task.

ROI selection

A region of interest (ROI) is a subset of a dataset particularly relevant for a specific purpose. In our study, a ROI is a part of a video frame that contains relevant information for our HR task. The proper selection of the ROI on a subject's face is crucial for efficient HR estimation. While several research works have analyzed facial regions for accurate estimation, these studies have primarily employed hand-crafted methods [114]. In Section 3.3.3, we contribute to this selection process by conducting an evaluation and comparison of the most commonly used ROIs for remote HR estimation.

Additionally, the size of the ROI is a critical consideration. Verkruysse et al. [48] studied the influence of ROI size on accurate rPPG signal extraction. They categorized ROIs into small, medium, and large windows. Consistent with their findings, we chose a region of medium size. Their study highlighted that small areas tend to result in more noisy predictions, as each pixel in the ROI has a significant impact on HR prediction. On the other hand, large zones are more sensitive to motion artifacts and may include non-skin pixels.

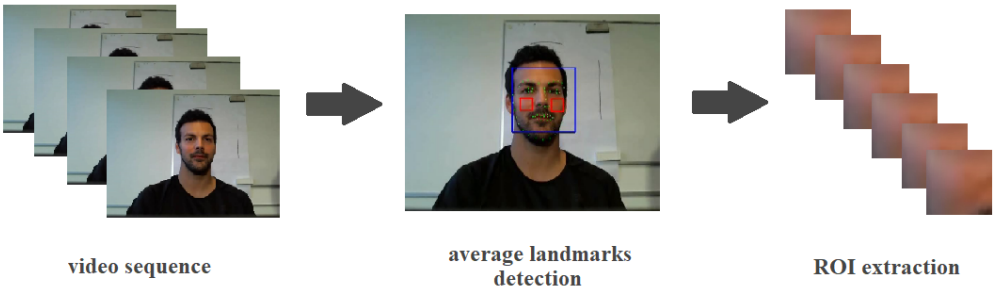


Figure 3.5: Diagram of the proposed ROI extraction approach from a video sequence for HR estimation.

We extract the cheek area from the face videos as described in Figure 3.5. First, we divide our videos into 5-second sequences of images creating sub-videos. We detect from every frame in the sub-video the location of the 68 (x,y)-coordinates of the dlib landmark detector to map the shape of the face on the image. Then, we obtain the average landmark points per sub-video and based on those, we compute a 40×40 pixels sized region of each of the two cheek areas for every frame of the sub-video. Most of the HR measurement methods tend to average the color values in the entire ROI and use them as the original rPPG signal. By performing this step, we lose the local information within each ROI, therefore, we choose to pass it entirely as input to our neural network.

Green channel selection

In early studies, the strength of the plethysmographic signal in the G channel of a face video was proved sufficent [48]. This is consistent with the fact that hemoglobin absorbs green light better than red and blue [115] light. However, Verkruysse *et al.* highlighted the fact that the Red (R) and Blue (B) channels may contain complementary information, this is why in Section 3.3.3 we perform an evaluation of the effectiveness of the G channel selection for our method compared to the choice of R and B channels and the use of the three RGB channels as originally provided in the video. Other deep learning approaches [66] used as input of their structures just the G channel of the face videos although our approach differs from theirs in a crucial point: they consider the selection of the G channel as a way to leverage the tasks of a CNN, reducing the number of parameters of the network without any study that justifies the selection.

Neural network

CNNs are a type of deep learning model that typically operates directly on raw inputs, such as images, to extract patterns for various tasks. CNNs have proven to be very efficient, especially for classification tasks, which are the focus of this paper. However, these models are often designed to handle 2D inputs. A three-dimensional CNN (3D-CNN) is a network of processing layers used to reduce three-dimensional data to its key features, making it more suitable for classification. We represent our input data in a three-dimensional format, where the first two dimensions correspond to the 2D images, while the third dimension represents time.

Recent works in the literature have demonstrated the success of 3D-CNN structures in handling 3D data, such as videos [116], [117]. We believe in the potential of 3D-CNNs for extracting rPPG information embedded in human faces, and it seems this type of network has not been fully explored yet. While two other works attempted HR estimation using 3D-CNN, major drawbacks in those approaches motivated us to propose an optimal and reproducible option. Bousefsaf *et al.* [66], presented a 3D-CNN that produces predictions for every pixel in a video stream, resulting in a heavy network with processing times of days for one test video. On the other hand, Hu *et al.* [70], performed a comparison between 2D-CNN and 3D-CNN. Although the rough implementation of 3D-CNN proved more suitable for HR estimation, the main focus of Hu *et al.* work was on adding modules to the 2D-CNN structure, leading to an underutilization of their 3D-CNN.

The architecture of the selected 3D-CNN is shown in Figure 3.6. The input video patch samples are of the size $(300, 40, 40, 1)$ being 300 the number of frames, $(40, 40)$ the ROI size defined and 1 representing the G channel. This input data is passed to the first *convolution layer*, where the video patch is transformed by kernels, sets of learnable filters. The *convolution layers* are followed by *pooling layers*, where filters evaluate small sections at a time to abstract the values to maps. We use *maxpooling layers*, which act as a noise suppressant by taking the highest value of an area. After an alternated use of *convolution layers* and *pooling layers*, our network has two *dense layers*, resulting from flattening the last *maxpooling layer*. Our last *dense layer* implements a *softmax function* which assigns decimal probabilities to each class to solve the multi-classification problem. Those decimal probabilities add up to 1 for faster convergence. We decided to exploit the softmax function at the output layer of our network as a way to handle outliers for a better estimation of the HR. By leveraging classification over regression, our network is more resilient to outliers.

When the probabilities are identified and analyzed, the output is assigned

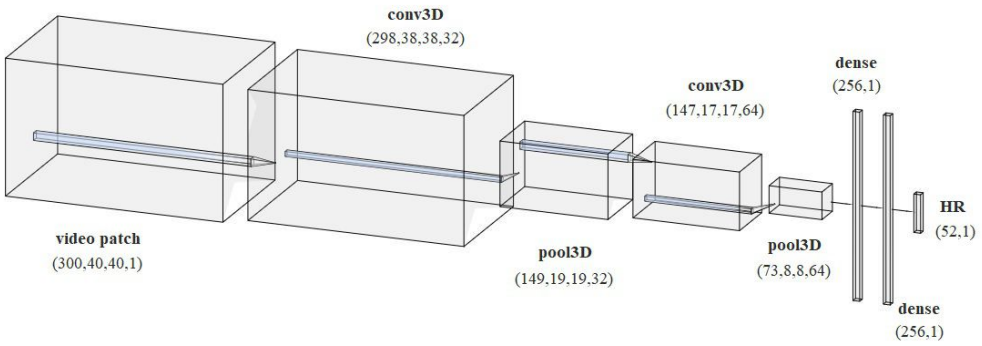


Figure 3.6: Proposed 3D-CNN architecture for HR estimation. The network takes the data as a 3D input, then alternates between 3D Convolutional layers and 3D MaxPool layers, ending with two fully connected layers that output the estimated HR.

to a value, in our case, a *one hot encoding* representation of the HR. "*One hot encoding*" is a process by which categorical variables are converted into vectors composed of the digit "1" only in the position of the class predicted with the highest probability and the digit "0" for the rest of the classes. The output of our network is then a vector of length l, being l the number of classes. In this case, $l = 52$ classes from 48 to 100 bpm with a step of one. Finally, after conversion from *one hot encoding* vectors to a scalar, we perform an average for all the predictions per sub-video for both cheeks, computing the final HR prediction.

3.3.2 Experimental Setup

Dataset

Many existing methods have reported their results using private, self-collected datasets, making it challenging to compare the performance of individual approaches. Our aim is to demonstrate the capability of our method to perform under different illumination conditions, particularly when parts of the subject's face are barely illuminated, such as when the light source is coming from one side rather than the front.

Additionally, we seek to confirm the robustness of the model against various challenges, including head pose variation, different skin tones, and facial expressions, creating a well-simulated and realistic scenario. To validate these hypotheses and enable a fair comparison with other approaches, we evaluated our method on the publicly available and challenging dataset,

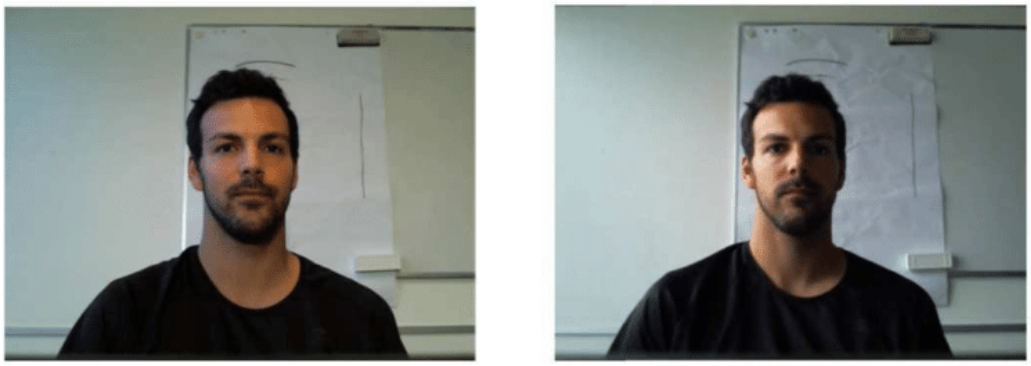


Figure 3.7: Example video frames of two videos from a subject of the COHFACE dataset. Frame (a) shows the subject's face illuminated with studio light and frame (b) with daylight coming from a left source.

COHFACE. The COHFACE dataset [49] is composed of 160 facial videos captured at 20 frames per second (fps) with a resolution of 640×480 collected from 40 healthy individuals and their physiological signals. The dataset includes 12 female and 28 male subjects between 19 and 67 years old. Each video has a length of 60 seconds. Physiological readings were taken by a BVP sensor, which measures changes in skin reflectance to near-infrared lighting caused by the varying oxygen level in the blood due to heart beating. We converted the BVP signals to HR measurement using a function from the Bob package `o`.db.cohface [118]. The videos in this dataset have realistic illumination conditions, the subjects are recorded under two different lighting conditions as shown in Figure 3.7: (a) Studio, closed blinds, avoiding natural light, and using extra light from a spot to homogeneously illuminate the subject's face, (b) Natural, all the lights turned off and open blinds. The daylight videos (b) pose a significant challenge in this research, as the right side of the subject's face is poorly illuminated, resulting in pixel values close to 0 for every channel. This creates dark ROI videos that may act as disturbances to the network during the learning process. However, as discussed by Hernandez *et al.* [119], a diverse training dataset representative of realistic conditions allows deep learning models to extract information independent of the acquisition scenario. Leveraging the self-learning characteristic of neural networks, we address the challenges presented in COHFACE. Similar to other works [54], we performed a subject-exclusive split of the videos for training and testing subsets. The training set is composed of 28 subjects and a testing set of the remaining 12.

Metrics

Different metrics have been used in the literature for reporting the HR estimation performance of an approach. Evaluating a deep learning algorithm with different evaluation metrics is an essential part of its validation because it gives an overall assessment of a model's performance. We present the Mean error of the predictions (Me) and its standard deviation (StD) in bpm of the HR error, the MAE in bpm, the Root Mean Squared HR Error (RMSE) in bpm, the Mean Absolute Percentage Error (MAPE), and Pearson's correlation coefficients.

Implementation details

The 3D-CNN structure was implemented in TensorFlow and Keras using a standard chain of conv3D layers, maxpool3D layers and activation functions. After each maxpool3D layer, a batch normalization was applied. Batch normalization was initialized with weights randomly sampled from a Gaussian and their values were scaled with a value γ and shifted with a value β, parameters learned during training. This was performed to avoid a linear activation of the inputs. A dropout of 0.5 was applied after each batch normalization to ensure a good training process by preventing model overfitting. Rectified linear activation functions were used in every conv3D and dense layer.

The size of the kernels was set to $3 \times 3 \times 3$ for the convolutional layers and to $2 \times 2 \times 2$ for the max pooling layers. The weights of the kernels were initialized and sampled from a normal distribution with a mean of zero and a standard deviation of $\sqrt{\frac{\alpha}{n}}$ with n equal to the number of input samples. The model was trained for 10 epochs, Adam optimizer was selected with a learning rate set to 0.001 and the loss function chosen was categorical cross-entropy.

3.3.3 Results

The results in Table 3.6 demonstrate that the proposed 3D-CNN structure exhibits competitive performance, achieving the lowest StD in the COHFACE dataset. It outperforms HR-CNN [54], confirming that the sequential processing of spatial and temporal information proposed by the authors cannot capture HR information as effectively as our network, which processes both spatial and temporal information simultaneously. Our model also surpasses 2 STREAM CNN [64] for every reported metric, indicating that the simultaneous 3D convolutions across all input video patches provide superior performance compared to the double-stream approach.

Furthermore, when using the cheek area of the video as input, our model

Table 3.6: Comparison between our proposed 3D-CNN and other AI-based approaches for HR estimation on the COHFACE dataset.

Method	**StD**	**MAE**	**RMSE**	**MER**	ρ
HR-CNN [54]	-	8.10	10.78	-	0.29
2-STREAM CNN [64]	-	8.09	9.96	-	0.40
3D-rPPGNet [70]	8.98	5.86	9.12	-	-
2D-rPPGNet [70]	8,08	5.59	8.12	-	**0.63**
AND-rPPG [69]	7.83	6.81	8.06	-	**0.63**
3D-CNN [C5]	**7.23**	**5.5**	**7.74**	**7.12**	0.62

achieves lower MAE and RMSE compared to the denoising patches obtained from the full face in AND-rPPG [69]. This highlights the superiority of selecting an optimal face region over denoising the entire face. Our model also outperforms, in almost every metric, the two networks that aimed to create a CNN-based feature maps extractor from full faces [70]. In their work, Hu *et al.* initially presented rough versions of 2D and 3D-CNN, their further improvements focused on the 2D model, letting aside the 3D model. In addition, the results highlight the optimal performance of our 3D-CNN: an end-to-end 3D CNN can outperform a 2D structure in accurately estimating a subject's HR directly from the cheek area without the need for any other intermediate face representation. This provides a new approach to capturing rPPG information without compromising model accuracy.

Furthermore, the processing time for a 60-second video at 20 fps with our 3D-CNN is only 0.1 milliseconds. The proposed 3D network requires no extra pre or post-processing steps, making it highly efficient and suitable for online estimation.

Effectiveness of input choice selection

We also conducted a study on the effectiveness of different video input choices, specifically ROI and input channel selection. As a baseline experiment, we trained and tested our 3D-CNN on full-face, three-channel videos. Subsequently, each of the RGB channels was used individually to train the network on full-face videos, and the results are reported in Table 3.7. The experiments suggest that, while there is no clear superiority between using RGB vs. G as input, selecting only the R or B channels significantly decreases the network's performance.

In the next step, we trained and tested the 3D-CNN by feeding it with a 40×40 and 30×80 ROI for the cheeks and forehead experiments, respectively. These areas were detected and cropped using the 68 (x,y)-coordinates of the

Table 3.7: Evaluation of the proposed 3D-CNN architecture using different input video channels and ROI on the COHFACE dataset.

Method	Me	StD	MAE	RMSE	MER	ρ
Full face RGB	2.43	9.55	8.22	9.86	11.32	0.28
Full face R	4.05	11.08	9.82	11.80	12.87	0.11
Full face G	1.44	10.35	8.23	10.45	11.54	0.29
Full face B	-0.29	10.47	8.95	10.47	12.79	0.23
Forehead RGB	-1.22	10.61	8.35	10.68	12.13	0.42
Forehead G	-3.17	8.80	7.84	9.35	11.71	0.52
Cheeks RGB	**0.01**	7.99	5.78	7.99	7.99	0.46
Cheeks G	2.75	**7.23**	**5.5**	**7.74**	**7.12**	**0.62**

dlib landmark detector, as explained in Section 3.3.1. The results presented in Table 3.7 prove that using a smaller and more specific area than the full face is particularly beneficial for achieving accurate HR estimation, especially in the case of the cheeks region. The cheek area is less affected by nonrigid motion, such as smiling or talking, and can yield better results. This is because, in some cases, the forehead can be occluded by hair or other monitoring devices. However, both regions can be equally affected by difficult illumination conditions. Notably, the results for both areas are especially promising for the G channel, highlighting the 3D-CNN's successful HR prediction even in adversarial illumination conditions (e.g., natural light sources that do not evenly distribute light on the face skin areas).

3.4 Summary

Traditionally, the primary goal of employing facial soft biometrics has been to enhance face recognition systems. To achieve this, extensive research has been dedicated to estimating soft biometric traits from faces, including gender, age, and ethnicity. However, nowadays the exploration of biometric feature estimation extends beyond the mere identification of individuals of interest in applications such as telehealth and daily parameter monitoring.

In this chapter, we extensively explore optimal deep training protocols for estimating two soft biometric traits—weight and heart rate. In our first study, we advance towards enhancing the accuracy of deep learning models for remote weight estimation from faces, achieving a 15.3% reduction in test error and thereby advancing the state-of-the-art on the public VIP attribute dataset. Introducing the new Prisoners dataset, characterized by a

diverse weight distribution compared to existing datasets, we conduct intra and cross-dataset experiments, evaluating the model's performance with traditional metrics and employing explainability techniques. Three critical factors in building automatic weight estimators from faces—gender, cropping bounding box, and facial hair occlusions—are presented and thoroughly assessed. Additionally, in a second study, we propose a competitive, fast, and reproducible heart rate estimation method based on a 3D-CNN structure. This method is evaluated against similar state-of-the-art deep learning structures on the publicly available dataset COHFACE. We address gaps in the current literature by performing key experiments, including a comparison of the most common ROI for remote heart rate estimation. We achieve optimal results using the cheek area and evaluate the choice of the G channel as input compared to using all three channels of RGB videos.

In the next chapters, we explore thermal imagery as a promising alternative to solidify different AI-based soft biometric estimators. This exploration serves as a solution to overcome the limitations posed by the visible spectrum and opens up new opportunities.

Chapter 4

Facial Soft Biometric Estimation in the Thermal Domain

Although thermal face recognition has recently become an active area of research, the estimation of other biometric traits from facial visuals in this spectrum remains largely unexplored. Additionally, there is a shortage of available thermal datasets designed for facial processing tasks, with existing ones being limited in terms of annotation. In this Chapter, a novel dual face dataset, acquired simultaneously in visible and thermal spectra and extensively annotated with soft and hidden biometric traits, is presented. Experiments are conducted on this dataset to explore the potential of thermal images for gender, age, and weight estimation.

Section 4.1 introduces thermal imagery and its traditional use in face biometrics. Following that, Section 4.2 describes existing biometric datasets containing thermal face visuals, and subsequently, Section 4.2.1 presents the LVT Face Dataset collected in the context of this thesis. Section 4.3 includes a description of the architectures implemented for visible and thermal soft biometric estimation as well as the evaluation protocol and metrics used in our experiments. In Section 4.4 the results of testing the networks on our LVT dataset are discussed. Finally, Section 4.5 outlines and summarizes the contributions made in this chapter.

The research question that this chapter aims to answer is: **RQ2**, i.e. *Are there benefits to extracting soft biometrics from thermal rather than visible data?*

4.1 Introduction

Traditionally, face processing models have based their estimation on images acquired in the visible spectrum. Despite the practical success and maturity of these networks, deep learning approaches based on visible spectrum images face challenges such as occlusion and illumination changes. Thermal imagery, on the other hand, has demonstrated remarkable success in facial recognition scenarios, being the base of the development of several effective models [120]. It has proved itself as a powerful caption tool [1] and has been presented as superior to visible imaging in hard conditions such as the presence of smoke, dust and absence of light sources [121].

Thermal imagery operates by detecting electromagnetic radiation in the medium MWIR ($3 - 8\mu m$) and long LWIR ($8 - 15\mu m$) wave infrared spectrum [122] where skin heat lays within. This capability enables thermal images to overcome the lack of illumination or some types of occlusions. However, works have highlighted how the thermal heat captured by thermal cameras can be affected by various factors such as ambient temperature or intense physical activity [1].

4.2 Existing Thermal Datasets

Interest in employing thermal face images has grown in the past years, nevertheless, this regard has been mostly confined to tasks such as landmarks and face detection and FR [1], [123]. A relevant subset of FR is Cross-FR (CFR) discipline that aims to identify a person's image in the thermal spectrum from a gallery containing face images acquired in the visible spectrum [120]. Only a few datasets have been provided involving visuals acquired in thermal spectra, and among them, those covering soft and hidden biometric metadata are few. In Table 4.1, we present an exhaustive selection of relevant datasets that include visuals in the thermal spectrum and some key descriptors of them including their year of release, the number of subjects, images and videos present in the dataset and their initial intended purpose.

One of the first datasets containing thermal visual data was presented in 2003 [124]. The data was acquired at the University of Notre Dame and contains images from 240 distinct subjects with four views showing different lighting and facial expressions with the purpose of recognizing individuals. Beyond people recognition, Wang *et al.* established a similar dataset for expression recognition, containing both spontaneous and intended expressions of more than 100 subjects [125], while Gault *et al.* recorded thermal videos from 32 subjects under three imaging scenarios and their paired rPPG

Table 4.1: Relevent face datasets containing visuals in thermal spectra.

Year	Dataset	# subjects	# images	# videos	Application
2003	UND-X1 [124]	241	4584	-	FR
2010	NVIE [125]	215	Unknown	Unknown	Expression recognition
2013	TH-HR [126]	32	-	96	HR
2018	VIS-TH [1]	50	2100	-	FR
2018	TUFTS [127]	113	10000	113	FR
2018	TH-HR-RR [128]	20	-	40	HR, RR
2021	Speaking faces [31]	142	-	45 hours	Biometric Authentication
2021	ARL-VTF [129]	395	549712	-	Cross-FR
2022	SF-TL54 [123]	142	2556	-	Landmarks detection
2023	LVT [C4]	52	612	416	Facial processing

signals for HR estimation [126]. In 2018, two new datasets were acquired for FR with multiple illuminations, pose, and occlusion variations [1], and including imagery from different modalities, namely visible, thermal, near-infrared, and a computerized facial sketch, and 3D images of each volunteer's face [127]. In the same year, Barbosa *et al.* collected thermal videos from 20 healthy subjects in two phases: phase A (frontal view acquisitions) and phase B (side view acquisitions), and the corresponding PPG and thoracic effort were simultaneously recorded for HR and Respiratory Rate (RR) estimation [128]. More recently, two large-scale visible and thermal datasets have been assembled. Abdrakhmanova *et al.* gathered a combination of thermal, visual, and audio data streams to support machine learning-based biometric applications [31], and Poster *et al.* presented the largest collection of paired visible and thermal face images to date. Variability in expression, pose, and eyewear were recorded [129]. Following, a thermal face dataset with annotated face bounding boxes and facial landmarks composed of 2556 images was introduced [123].

4.2.1 LVT Face Dataset

To the best of our knowledge, this section introduces the first dataset composed of visible-thermal paired face images and recordings along with associated information os the subject, namely gender, age, body temperature, SpO2, BP, HR (resting and after physical activity), height, weight, BMI and 11 additional health metrics. The comprehensive annotation of numerous parameters for each subject aims to facilitate the exploration of thermal data's potential

in assessing an individual's health status. Furthermore, it aims to enhance the accuracy of estimating soft biometric traits such as gender and age from images within this specific spectrum.

Acquisition material: The visible and thermal face visual data were obtained using the dual sensor from the FLIR Duo R camera developed by FLIR Systems. This camera is specifically designed to capture visible and thermal visuals simultaneously, particularly suitable for unmanned aerial vehicles. The FLIR Duo R dual camera has been employed in recent research due to its appropriateness in data collection for various tasks such as face recognition and cross-spectrum applications [1], [121]. The visible and thermal sensors of this camera consist of a CCD sensor with a pixel resolution of 1920×1080 and an uncooled VOx microbolometer with a pixel resolution of 640×512, respectively. To assess the health status of the subjects, various devices were utilized. A contactless infrared thermometer with a precision of ±0.2°C between 34°C and 42.0°C and a precision of ±0.3°C in the range of 42.1°C and 43.0°C was employed for computing the user's body temperature. For calculating BP, an OMRON HEM-7155-E tensiometer was used, along with a LED finger oximeter for SpO2 measurement with a precision of ±2%. To track HR, subjects wore a Garmin Vivoactive®4 smartwatch equipped with an optical PPG sensor capable of detecting the heart rate by shining a green light through the subject's skin, reflecting the red cells in the skin's blood vessels. For quantifying bodyweight-related measures, the RENPHO®Body Fat Smart scale was utilized. When a subject steps on the device and enters their gender, age, and height into the system, the scale returns 13 metrics, including weight and BMI.

Visuals collection protocol: The image and video acquisition took place in an indoor environment with the ambient temperature set to 25°C. In Figure 4.1, we present the arrangement. The acquisition setup included a white wall serving as a background, and a chair positioned at a fixed distance of 0.25 m from the camera, which was placed at a height of 1 meter from the ground. Additionally, a two-point lighting kit was strategically placed to minimize shadows, facilitating the segmentation of the subject from the background. Each volunteer participated in two separate acquisition sessions, with an average time interval of 6 weeks. Before the acquisition process, volunteers were requested to fill out and sign consent forms. The visual data comprises 6 images per person (3 visible and their associated thermal pair) in each session, encompassing three different conditions: Neutral (N), Ambient light (A), and an occlusion in the form of eyeglasses (O), resulting in a total of 612 images. Figure 4.2 illustrates example images of different individuals across different variations from the LVT dataset. Additionally, four 60-second videos were recorded per subject in each session under N conditions. The first

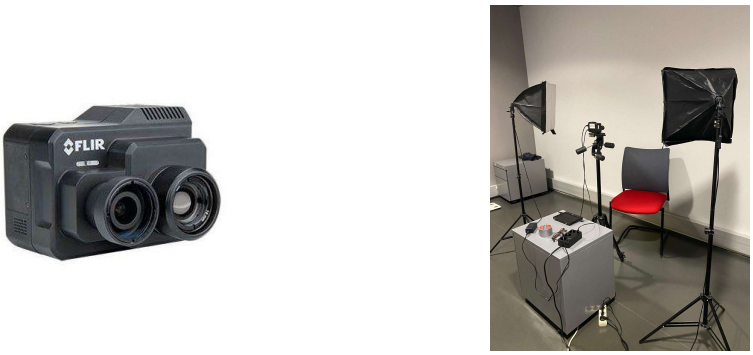


Figure 4.1: The material for the LVT Face Dataset collection is presented: Flir Duo R camera (left) and acquisition setup (right).

Figure 4.2: Example images from the LVT Face Dataset. The three variations are displayed in visible (upper row) and thermal (bottom row) spectra, from left to right: N, O and A.

pair of videos (one in the visible spectrum and its paired thermal counterpart) was captured after the subject had been resting for at least 5 minutes, while the second pair followed moderate exercise in the form of climbing stairs to elevate their HR values, resulting in a total of 408 60-second videos.

Subjects' metadata: Several pieces of metadata were collected to describe the subjects, including gender, age, and height. Additional parameters were quantified to assess their health status, such as body temperature, HR, BP, SpO2, weight, and BMI. In addition to weight and BMI, the smart scale provided 11 other variables: body fat and body water percentages, skeletal muscle, fat-free weight, muscle mass, bone mass, protein, subcutaneous and visceral fat, Basal Metabolic Rate (BMR), and metabolic age. The filenames

Table 4.2: (Table is read horizontally) Summary of the visuals and metadata contained in the LVT Face Dataset.

Identities	52 subjects	2 sessions

Visuals	6 paired images in conditions N, O A	2 paired 60s videos resting	2 paired 60s videos after activity

Metadata	**Biometrics**	ID	Gender	Age	Height	Weight
	Health paramet.	Temperature	HR resting	HR activity	BP max	BP min
		BMI	Body fat(%)	Body water(%)	Muscle	Fat-free kg
		Body mass	Bone mass	Proteins	Subcutaneous fat	
		BMR	SpO2	Visceral fat	Metabolic age	

for images and videos are constructed by indicating the visual data spectrum, subject ID, session ID (1 or 2), and in the case of images, the conditions at the time of acquisition (N, O, or A).

Summary: The LVT Face Dataset [D2] is designed as a collection of images, videos, soft biometrics, and health parameters recorded from 52 different subjects across two sessions. It consists of 612 face and shoulders images and 416 60-second videos, totaling approximately 285 GB of disk space. The 52 recorded participants, comprising 38 males and 14 females, are from 13 different countries spanning four continents, with ages ranging between 22 and 51 years. Out of the 52 subjects, 50 were present for 2 sessions, while 2 attended only one session. An executive summary of the dataset is provided in Table 4.2.

4.3 Methodology

In this section, we describe the models implemented and compared in our experiments, the evaluation metrics, and the experimental setup of the networks.

4.3.1 Soft Biometric Estimation Models

VGGNet [130] was developed by the Visual Geometry Group from the University of Oxford with the objective of improving computer vision tasks by increasing the depth of an architecture with very small convolutional filters of size 3×3. In addition to this, VGG incorporates 1×1 convolutional layers to make the decision function more non-linear without changing the receptive fields. VGG architecture has been proven in the literature as powerful for

estimating gender and age from face images [131]. Moreover, in their comparative study of architectures for gender estimation from thermal data, Farooq *et al.* [30] revealed the high performance of VGGNet for this task. No study, to the authors' knowledge, has been conducted on the feasibility of thermal imagery for age estimation. Therefore, we select the VGGNet network with 16 weight layers, i.e., the VGG16 model, for our gender and age estimation networks. We use the VGG16 base architecture as a feature extractor and we add custom fully connected layers on top for binary classification and regression for gender and age prediction respectively.

ResNet [132] are convolutional neural networks that introduce the concept of residual learning. Instead of learning a direct mapping between layers, ResNet learns the differences between the input and the desired output of a layer by using shortcut connections, also known as skip connections that bypass one or more layers and directly connect the input of a layer to its output. As presented in Section 3, face image-based weight estimation has been performed in this thesis via a ResNet architecture with 50 layers and a final regression layer. In this Chapter, we test the suitability of ResNet50 for weight estimation in the thermal spectrum.

4.3.2 Experimental Setup

Evaluation Metrics

Accuracy is used as a metric for gender classifier assessment. Regarding age and weight, we report the MAE and the MSRE in years and kg respectively and the Pearson's correlation coefficient. Additionally, for age, we provide the StD of the difference between the predicted and the real age of the subjects. Finally, we include the PAP for the weight estimation network.

Implementation Details

The VGG16 and ResNet50 architectures were implemented using TensorFlow and Keras frameworks. The weights of the VGG16 model were initialized with pre-trained weights obtained from the ImageNet dataset with its final fully connected layers excluded to add custom layers designed for the specific tasks under consideration. The output of the VGG16 base model was flattened and passed through a fully connected layer consisting of 256 neurons with ReLU activation. Additionally, a dropout of 0.5 was applied to prevent overfitting. Finally, for the binary gender classification task, a final output layer with a single neuron and sigmoid activation function was added, while for the age regression task, an output layer with a single neuron and linear

Figure 4.3: Transfer learning protocol for soft biometric estimation from visible and thermal images.

activation function was incorporated. For the ResNet50 model, the weights were initialized with pre-trained weights obtained from the UTK dataset.

The input images were resized to 224 × 224 pixels. Each network underwent two training sessions with identical configurations: one with visible images and another with thermal images, utilizing Transfer Learning (TL). A subject-exclusive split of the dataset was conducted, allocating 240 images for training and 240 for testing. The training and testing pipeline are illustrated in Figure 4.3.

The VGG16 networks were trained for 20 epochs using the Adam optimizer with a learning rate of 0.001. Binary cross-entropy loss function was selected for gender classification, while mean squared error was employed for age estimation. Each ResNet50 model was re-trained during 10 epochs followed by an additional 10 epochs for training the final regression layer. During each TL step, the first 20 layers were frozen. Adam optimizer was used with a

Table 4.3: Performance of the selected VGG-16 network for gender classification given thermal and visible face images from the LVT test set.

GENDER	**VIS**			**TH**		
	N	**O**	**A**	**N**	**O**	**A**
Accuracy	**0.85**	**0.80**	0.75	0.80	0.75	**0.80**

Table 4.4: Performance of the selected VGG-16 network for age estimation given thermal and visible face images from the LVT test set.

AGE	**VIS**			**TH**		
	N	**O**	**A**	**N**	**O**	**A**
Std	7.08	7.24	6.82	**6.47**	**6.48**	**6.42**
MAE	5.77	6.20	5.53	**3.95**	**4.08**	**3.80**
MSRE	8.73	9.18	8.57	**7.31**	**7.44**	**7.25**
ρ	0.29	**0.32**	0.24	**0.35**	0.31	**0.39**

learning rate of 0.01, and Huber loss function was selected with $\delta = 1$.

4.4 Experimental Results

In this Section, we present a comparative study of state-of-the-art networks for estimating soft biometrics using visible and thermal images as input data. Additionally, we compare thermal and visible domains across various facial variations introduced in our dataset, assessing the performance of both modalities in practical scenarios.

Research has demonstrated that bone, muscle, and body fat exhibit unequal thermal conductivity [133]. Heat emission patterns can be used to characterize individuals by providing insights into major blood vessel locations, skeleton thickness, tissue amounts, and muscle and fat distribution[1]. Moreover, as proved in Chapter 3, male and female subjects have differing facial appearances even when of the same weight [32]. Thus, we infer that thermal imagery can offer vital information for the various soft biometric tasks explored in this research.

In Table 4.3, we present the results of the VGG16 network trained on the visible (VIS) and thermal (TH) images of our dataset for gender classification. We observe that under N conditions, where studio lights are used, visible data outperforms thermal imagery. This trend continues when occlusions in the form of eyeglasses are present, as we see in O conditions. In the thermal

[1]https://biometrics.mainguet.org/types/face.htm#thermogram

Table 4.5: Performance of the selected ResNet50 network for weight estimation given thermal and visible face images from the LVT test set.

WEIGHT	**VIS**			**TH**		
	N	**O**	**A**	**N**	**O**	**A**
MAE	8.60	**9.01**	8.64	**6.84**	12.35	**7.05**
MSRE	11.01	**11.48**	10.86	**9.46**	14.91	**9.80**
ρ	0.33	0.38	0.38	**0.41**	**0.51**	**0.42**
PAP	40 %	**40 %**	40 %	**55 %**	20 %	**60 %**

spectrum, glasses act as opaque barriers, leading to a loss of information in the eyes and penalizing this modality. However, in scenarios where no dedicated light sources are applied such as A condition, VGG16 trained with thermal data delivers better results. Table 4.4 displays the results of the VGG16 networks trained for age estimation. In contrast to gender estimation, thermal imagery exhibits clear superiority across all data variations. Even in the presence of eyeglasses, the error is only slightly higher compared to other variations. Finally, Table 4.5 presents the results of the weight estimation network. Similar to gender estimation, the metrics demonstrate that ResNet50 performs better in weight estimation when using thermal data, particularly in N and A conditions. This confirms the potential of thermal imagery in capturing hidden and detailed information from human faces, especially for age and weight estimation tasks using VGG and ResNet50 architectures, respectively.

4.5 Summary

Although the estimation of soft and hidden biometrics from face visuals has grown as a major area of research in the past years, deep-learning-based models are still challenged by the lack of robustness in RGB, for instance, with changing illumination conditions.

Motivated by this, in this Chapter, we have presented the collected Label-EURECOM Visible and Thermal Face Dataset for face biometrics. This dataset is the first that contains paired visible and thermal images and videos from 52 subjects with metadata of 22 soft biometrics and health parameters. In addition, we present the first comparative study between visible and thermal spectra as input images for soft biometric estimation, namely gender age and weight, from face images on our collected dataset. The experiments conducted demonstrate the feasibility of accurately estimating these three biometric traits from facial thermal data. To provide a comprehensive analysis, we

partition the test set into three subsets based on the three variabilities present in the LVT dataset: studio lights, occlusion in the form of eyeglasses, and ambient light. Through this approach, we establish the superiority of thermal imagery, particularly in age and weight estimation from faces. Furthermore, we prove that the performance of thermal imagery is superior to visible when no dedicated light sources are employed, such as in ambient light conditions.

Building on these promising results, Chapter 5 explores thermal data not only as an alternative but also as a complement to visible data through fusion techniques. Additionally, a novel architecture, previously unexplored for facial processing tasks, is trained and compared against the traditional networks presented in this Chapter.

Chapter 5

Fusion of Visible-Thermal Spectra for Soft Biometrics Estimation via Universal Face Embeddings

Vision Transformers have emerged as a powerful deep learning architecture capable of computing target outputs using multi-attention mechanisms that capture complex relationships between distant parts in input images. However, these models have not yet been applied to facial processing tasks. In this chapter, we introduce Bidirectional Encoder Face representation from Image Transformers (BEFiT), the first model that leverages the Transformer architecture to capture both local and global face features, resulting in a universal face embedding from which different biometric traits can be estimated. Building on the insights obtained in Chapter 4, we investigate the utilization of both visible and thermal imagery, as well as the fusion of both spectra, to estimate three different soft biometrics: gender, age, and weight.

This Chapter starts with Section 5.1, where the use of Transformers for facial processing tasks is motivated. Section 5.2 outlines the methodology employed for extracting the universal embedding using BEFiT, the fusion protocol, and the experimental setup while in Section 5.3 we present the performance analysis of our approach, which includes the utilization of both visible and thermal imagery and the fusion of scores from both networks for the estimation of gender, age, and weight. Finally, we conclude in Section 5.4 by highlighting the most important key points of the Chapter.

The research questions that this chapter aims to answer are: **RQ1**, i.e. *What are the most suitable deep learning architectures and training methodologies for various facial processing tasks?* and **RQ2**, i.e. *Are there benefits to*

extracting soft biometrics from thermal rather than visible data?

5.1 Introduction

Transformer models [134] have boosted the performance of deep learning models across various domains in the last years. Traditionally employed in Natural Language Processing (NLP) tasks, attention-based neural networks such as the Vision Transformers (ViTs) are now making significant progress in image-based tasks attaining state-of-the-art results on many computer vision benchmarks [135]. Vision Transformer [136] excel in image classification [137], object detection [138] and text-to-video translation [139]. Regarding facial processing tasks, two works have explored ViTs in the context of FR [140], [141].

Since their creation in 2016, Transformers [134] have proven superior to other similar structures such as Recurrent Neural Networks (RNNs) due to their ability to process data in parallel rather than sequentially. By leveraging self-attention mechanisms, Transformers can effectively capture relationships between different parts of input sequences, providing context that might not be discernible through sequential processing as the most relevant image patches for prediction may not necessarily be adjacent to the current one. This allows Transformers to process multiple sequences in parallel, speeding up the process thanks to the parallelization of attention mechanisms.

While CNNs have achieved remarkable success in facial processing tasks, they face a fundamental challenge in capturing long-range relationships among different facial regions. To capture long-distance dependencies, the traditional convolution model should enlarge its receptive fields through the stacking of convolutional layers. However, Vision Transformers offers a natural solution to this problem by learning global token dependencies within images [141].

5.2 Methodology

In this section, we describe the basic architecture of vision transformers, which consist of multi-head attention and feed-forward neural networks. Following that, we present our ViT proposal, BEFiT, as well as the processing that follows the face embedding computation, to estimate the desired soft biometric traits from different spectra.

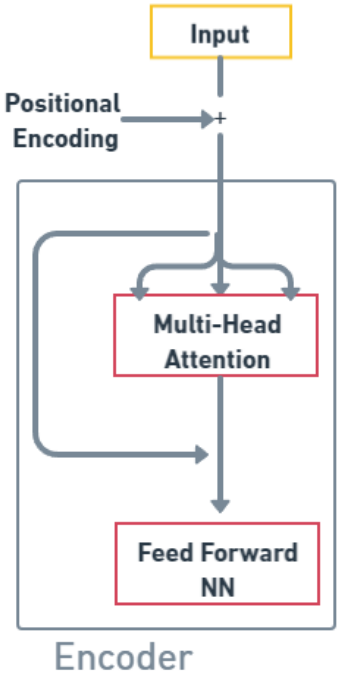


Figure 5.1: Vision Transformer Encoder structure

5.2.1 BEFiT Model

Vision Transformers [136] are a type of deep learning model that extends the Transformer [134] architecture, originally designed for natural language processing tasks, to handle computer vision tasks such as image classification, object detection, and segmentation.

In Figure 5.1 we present the basic architecture of vision transformers. To be processed by the vision transformer, each image is divided into fixed-size patches. Positional encodings are added to the patches to provide spatial information about the position of each patch in the image. Transformers operate through sequence-to-sequence learning, where the transformer takes a sequence of tokens (in our case, image patches) and predicts the next element in the output sequence. This process iterates through the encoder layers, with each layer generating encodings that define the relevance of each part of the input sequence to others, which are then passed to the next encoder layer.

The main advantage of transformers is the self-attention mechanism. The patch embeddings, along with their positional encodings, are fed into the self-attention mechanism where each patch embedding attends to all other patch embeddings, including itself, to compute a weighted sum representation of the entire image. In addition, vision transformers employ multi-head

attention, where the self-attention mechanism is performed multiple times in parallel with different sets of learned parameters. This allows the model to attend to different aspects of the input image simultaneously and learn diverse spatial relationships. After the self-attention mechanism, the output is passed through position-wise Feed Forward Networks (FFNs). FFNs consist of two fully connected layers with a non-linear ReLU activation function applied in between. These layers help capture spatial features within individual patches, allowing the model to encode local information in the image such as edges, textures, and shapes.

BEFiT is the first trained Vision Transformer that produces a general face embedding given a face image. Its architecture is based on the architecture of BEiT (Bidirectional Encoder representation from Image Transformers) [137] for image classification tasks. BEiT enhances the performance of other vision transformers by introducing a masked image modeling task as a pre-training step in the learning of the transformer. In computer vision tasks like image classification where the goal is to predict a single output based on the input image, only the encoder is required. After the pre-training step of the ViT for masked image modeling, we discard the decoder obtaining our pre-trained BEFiT. BEFiT is then trained for face recognition (FR). Our objective is to encode all the information that defines a person into a single vector, providing a meaningful representation from which different traits can be estimated without the need for costly re-training of the transformer every time a new facial processing task needs to be performed. Similar to how humans recognize each other in nature, we believe that by training the model for face recognition, other information—such as gender, age, nose size, eye color, etc.—will be embedded in the feature vector, as all these characteristics form part of one's identity. As depicted in Figure 5.2, after the training of BEFiT, given the input image patches, the model produces a fixed-size embedding representing the entire image. This embedding is then fed into a classification head to make predictions.

Figure 5.3 depicts the pipeline for soft biometric estimation. We train BEFiT for FR on RGB face images. After that, BEFiT is fine-tuned with thermal faces to perform FR in thermal spectra. We will refer to the different versions of BEFiT as BEFiT-V and BEFiT-T depending on the spectra in which they work. Afterward, customized CNNs are defined. These networks take as input the embeddings obtained with BEFiT and are trained to classify the three soft biometric traits studied in this chapter. For binary traits such as gender, we define a CNN consisting of a dense layer with 64 units and ReLU activation followed by an output layer with 1 unit and sigmoid activation for binary classification. For regression traits, the CNNs architecture consists of a sequential stack of two fully connected (dense) layers: the first layer has 64

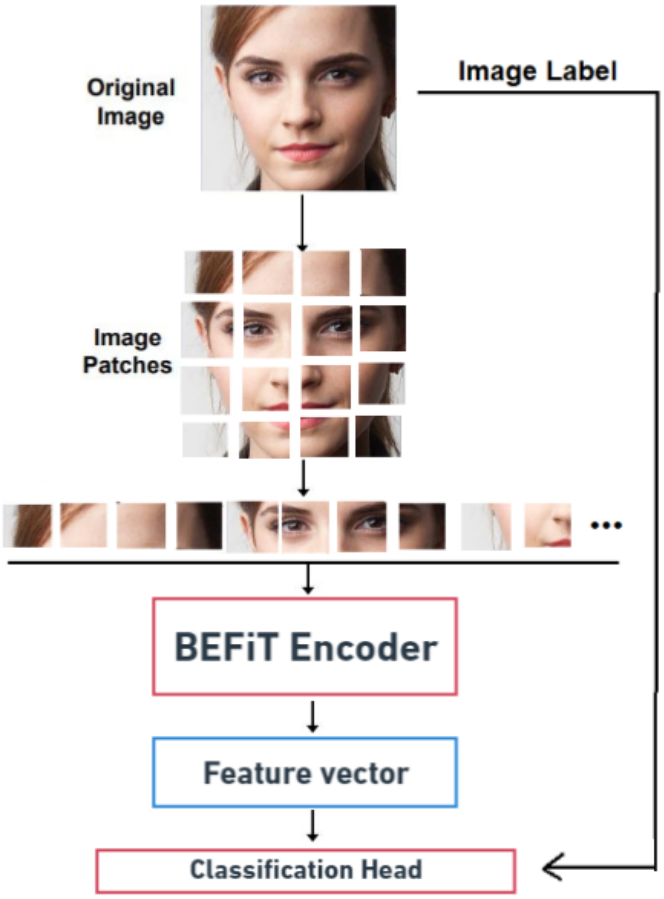


Figure 5.2: Overview of BEFiT inputs and outputs. Given an input image divided into patches, BEFiT produces a universal feature vector from which different facial traits can be inferred.

units and uses ReLU activation function, and the second layer has a single unit with ReLU activation function, which outputs the predictions. Fusion is performed at the decision level. It is applied to scores for classification tasks and to predictions for regression tasks. A weighted average is computed in each case.

5.2.2 Experimental Setup

Datasets

The training of BEFiT-V is performed using the CelebA dataset [142]. The CelebFaces Attributes Dataset (CelebA) is a large-scale face attributes dataset

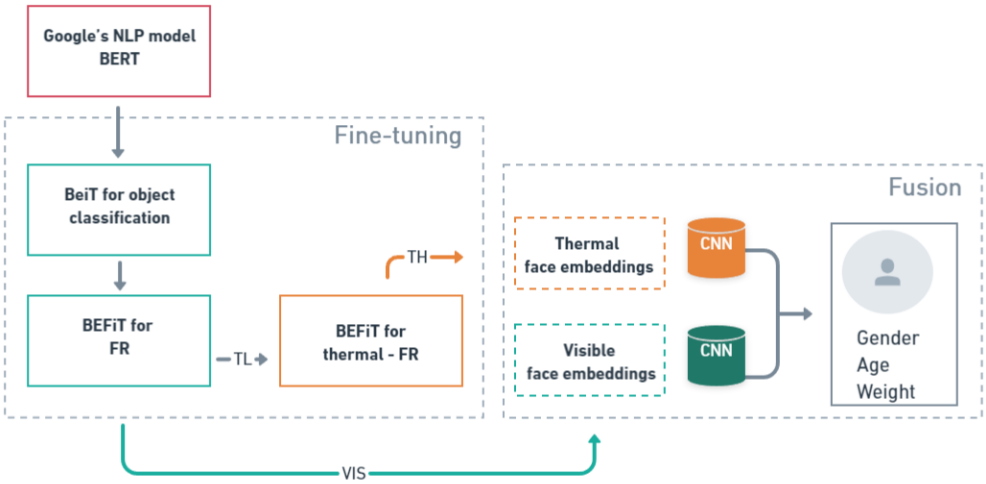


Figure 5.3: Fine-tuning pipeline for soft biometric estimation using BEFiT. Given any input face image, BEFiT-V and BEFiT-T compute a general face embedding. These embeddings serve as the foundation for estimating three key soft biometric traits: gender, age, and weight.

with more than 200K celebrity images from more than 10K unique identities. The images in this dataset cover large pose variations and background clutter. The CelebA dataset is traditionally employed as a training set for face attribute recognition, face recognition, face detection, and landmark localization among others.

BEFiT-T is fine-tuned on the TUFTS dataset [127], which was presented in 2018. The TUFTS dataset is composed of more than 10K images, including imagery from different modalities, namely visible, thermal, near-infrared, computerized facial sketch, and 3D images of each volunteer's face. This dataset has been employed for face recognition tasks.

As described in Section 5.2.1, custom CNNs are defined to estimate each task from the BEFiT embeddings. Different datasets are used to train the visible CNNs. The gender CNN is trained using the CALFW dataset [143]. The Cross-Age Labelled Faces in the Wild (CALFW) is an improved version of the LFW face dataset by adding face pairs with age gaps to incorporate the aging process intra-class variance while maintaining the same identities as in the LFW dataset. The CALFW dataset contains 4,025 individuals with 2, 3, or 4 images for each person and is used for face verification and gender classification tasks. We use the AgeDB [144] dataset to train the age CNN. AgeDB is a manually collected dataset with a wide range of ages for each subject, comprising 568 identities with 29 images per subject, resulting in

a total of 16,488 images. The weight CNN is trained on the VIP attribute dataset [39] described in Chapter 3, which is the largest face dataset annotated by weight.

As presented in Chapter 4, to the authors' knowledge, one dataset exists with paired visible-thermal facial images annotated with gender, age, and weight: The LVT Face Dataset for face biometrics. To provide a fair comparison between the visible and thermal networks, we perform the same subject-exclusive split of the LVT dataset in Training set (480 images from 40 subjects) and Testing set (120 images from the remaining 12 subjects) as the one defined in Chapter 4. The thermal gender, age, and weight CNNs are trained on the LVT training set. All the models are tested on the LVT test set.

For our cross-dataset experiments, we tested the architecture on the VIS-TH dataset [1]. This dataset consists of 2100 paired visible-thermal images captured under challenging conditions, including variations in expressions, head poses, occlusions, and different illuminations. It encompasses 50 subjects of diverse age, sex, and ethnicity.

Metrics

Accuracy is used as a metric for gender classifier assessment. Regarding age and weight, we report the MAE and MSRE in years and kg kg respectively and the correlation coefficient ρ. Additionally, for age, we provide the StD of the difference between the predicted and the real age of the subjects. Finally, we include the PAP for the weight estimation network.

Baselines

The performance of BEFiT for the estimation of soft biometrics is compared against the visible and thermal baselines presented in Chapter 4: VGG16 network for gender and age and ResNet50 for weight estimation. In addition, we test three publicly available and largely trained SotA networks. Those networks estimate the different soft biometric traits from images in the visible domain. For gender classification, we adopt the open-source *DeepFace*[1] library. Deep EXpectation for age estimation, (DEX)[2] [23] model is used for age estimation. Finally, for weight estimation, we use the ResNet50 implementation proposed in Chapter 3.

[1]https://github.com/serengil/deepface
[2]https://github.com/siriusdemon/pytorch-DEX

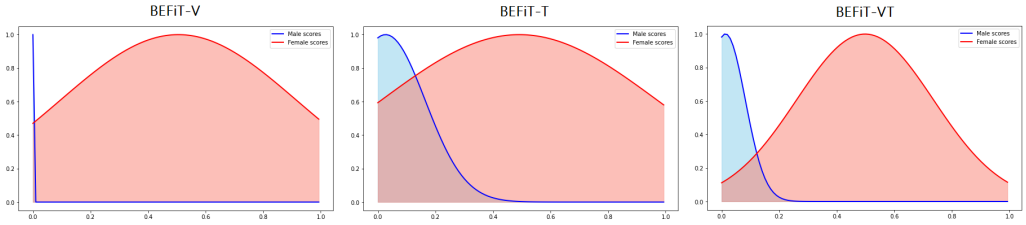


Figure 5.4: Scores distribution for male and female classes for gender classification via BEFiT model.

Table 5.1: Evaluation of the gender classification models in the LVT test set for different input data modalities.

GENDER	**Visible**			**Thermal**		**Fusion**	
	Deepface	**VGG16**	**BEFiT-V**	**VGG16**	**BEFiT-T**	**VGG16**	**BEFiT-VT**
Accuracy	0.79	0.81	0.95	0.77	0.86	0.82	**0.97**

Implementation Details

We initialize the training of BEFiT-V for face recognition with BEiT pre-trained values obtained from HuggingFace[3]. BEiT was pre-trained on ImageNet-22k, a collection of 14 million images and 22K classes. Each 224 × 224 image is divided into fixed-size patches of size 16 × 16. BEFiT-V and BEFiT-T were trained for 150 epochs in the CelebA and TUFTS datasets respectively with a batch size of 32, a learning rate of 0.002 and weight decay set to 0.05.

For soft biometric estimation, once the embedding was extracted, the CNN architectures were implemented using TensorFlow and Keras frameworks. The models were trained for 20 epochs, the parameters selected were a batch size equal to 32, Adam optimizer and a learning rate of 0.001. The loss function chosen was binary cross-entropy for gender classification and MAE for age and weight regression.

Fusion was performed at the decision level. A weighted average was calculated using values of 0.5, 0.2, and 0.7 for the output of the visible network corresponding to gender, age, and weight, respectively.

5.3 Experimental Results

In Table 5.1, we present the accuracy of the different approaches for gender classification. When comparing VGG16 and BEFiT for both thermal and visible spectra, we observe an advantage of using RGB images for predicting

[3]https://huggingface.co/docs/transformers/model_doc/beit

Table 5.2: Evaluation of the age estimation models in the LVT test set for different input data modalities.

AGE	**Visible**			**Thermal**		**Fusion**	
	DEX [23]	**VGG16**	**BEFiT-V**	**VGG16**	**BEFiT-T**	**VGG16**	**BEFiT-VT**
StD	8.69	7.04	9.50	6.45	5.56	6.50	**5.21**
MAE	7.23	5.83	8.41	3.94	4.36	4.11	**3.69**
MSRE	9.12	8.82	10.70	7.33	6.33	7.45	**5.40**
Correlation	0.53	0.28	0.31	0.34	0.45	0.32	**0.55**

Table 5.3: Evaluation of the weight estimation models in the LVT test set for different input data modalities.

WEIGHT	**Visible**			**Thermal**		**Fusion**	
	ResNet50[C3]	**ResNet50**	**BEFiT-V**	**ResNet50**	**BEFiT-T**	**ResNet50**	**BEFiT-VT**
MAE	**8.13**	10.11	11.29	8.18	11.12	9.11	9.16
MSRE	11.26	12.97	13.48	12.36	16.03	**10.18**	10.76
Correlation	0.57	0.39	0.31	0.76	0.18	**0.74**	0.37
PAP	53%	35%	33%	36%	**60%**	36%	**60%**

this trait. However, the superiority of BEFiT for extracting gender is clear, with BEFiT-V correctly classifying 95% of the subjects in the LVT test set. Moreover, by fusing the scores provided by BEFiT-V and BEFiT-T, BEFiT-VT achieves 97% correct classification. In Figure 5.4, the distributions of male and female scores obtained with the different versions of BEFiT are presented. In our training, the male class was set to zero and the female to one. We can observe that the female class has scores spread along the entire interval, while the male class reports scores very close to zero, especially remarkable in the case of BEFiT-V. Fusing both spectra allows for greater separability between the classes, consistent with the results of Table 5.1, where the accuracy of BEFiT-VT for the classification task is higher.

Table 5.2 presents various metrics assessing the different age estimators. Contrary to the results for gender classification, thermal imagery surpasses RGB for both architectures, VGG16 and BEFiT. The fusion strategy seems especially beneficial for BEFiT-VT, able to gather information from both spectra achieving the lowest errors in the LVT test set. The MAE of BEFiT-VT, at 3.69, is half that of the one presented by the SotA estimator DEX. In contrast to this behavior, it can be observed that in the case of VGG16, thermal predictions are generally penalized by their visible counterpart, resulting in less accurate results than using them alone.

In Table 5.3, the results of the comparative study of different techniques for weight estimation are displayed. In this case, the results indicate that training a dedicated network delivers more accurate results than extracting

Table 5.4: Evaluation of the gender and age estimation models in the VIS-TH dataset.

GENDER	**Visible**			**Thermal**		**Fusion**	
	Deepface	**VGG16**	**BEFiT-V**	**VGG16**	**BEFiT-T**	**VGG16**	**BEFiT-VT**
Accuracy	0.84	0.60	0.93	0.33	0.87	0.28	**0.97**
AGE	**Visible**			**Thermal**		**Fusion**	
	DEX [23]	**VGG16**	**BEFiT-V**	**VGG16**	**BEFiT-T**	**VGG16**	**BEFiT-VT**
StD	5.87	6.64	9.20	5.21	6.88	**5.19**	5.89
MAE	4.95	5.16	7.71	4.66	8.75	**4.55**	5.42
MSRE	6.17	6.67	9.44	5.49	10.73	**5.40**	6.89
Correlation	**0.47**	0.06	0.40	0.03	0.15	0.06	0.39

weight from a general face embedding. The superiority of thermal data is also confirmed for this task. When fusing the decisions in BEFiT-VT, the network achieves competitive performance with the state-of-the-art ResNet50. Indeed, BEFiT-VT has the lowest PAP in the LVT test set and competitive results in terms of MAE and RSME. The fusion strategy is also optimal for the ResNet50 networks, achieving the lowest MSRE and higher correlation coefficient in the LVT test set.

To assess the generalization of BEFiT, we have tested gender and age estimation on a more challenging dataset, because of face pose, expression, and illumination variation. Table 5.4 presents the performance of BEFiT and the SotA methods on the VIS-TH dataset. Again, BEFiT-VT performs best for gender estimation despite the more complex conditions whereas VGG has a big drop in performance.

As for age, by observing the results, we can confirm that thermal data has an advantage over visible data. Moreover, the fusion strategy proves to be the most successful once more for both architectures, BEFiT and VGG16. BEFiT-VT and DEX have similar performances, with a drop compared with the results on LVT. VGG16 achieves the best results in terms of MAE and MSRE, however, the low correlation coefficient obtained in each spectrum for the VGG16 architecture reflects that the predictions given by this architecture are always close to the dataset's average age, resulting in minimized age error without learning specific face features for age estimation.

5.4 Summary

Previous work on soft biometric estimation requires specialized networks per soft biometric trait to be estimated. As an alternative, approaches such as

multi-task learning are proposed, but their performance comes at the cost of network complexity. In addition, many soft biometric traits can be estimated from the face; consequently, multiple training sessions need to be done.

In this Chapter, we have introduced a novel structure for face embedding extraction: BEFiT. BEFiT is a vision transformer that can extract a unique face embedding from which different soft biometric traits can be estimated. Unlike other approaches for soft biometric estimation, the training of BEFiT for face feature vector extraction was not optimized for a specific soft biometric trait estimation, thus boosting embedding generalization.

We train two different versions of BEFiT (BEFiT-V and BEFiT-T) in the visible and thermal spectra, and we compare their performance with state-of-the-art networks and baselines demonstrating that the BEFiT embeddings can capture essential information for gender, age, and weight estimation, surpassing the performance of dedicated deep learning structures for each soft biometric trait. Additionally, we fuse BEFiT-V and BEFiT-T at the decision level enhancing the performance of the soft biometric estimators by gaining insights from both visible and thermal spectra.

Part II

Influence of Social Media Filters in Facial Biometrics

Chapter 6

Impact of Beautification Filters in Face Biometrics

This Chapter introduces pioneering research on evaluating the influence of social media filters, specifically those designed for beautifying individuals, in AI-based face models. The investigation targets networks SotA in security contexts, particularly within face verification scenarios and deepfake detection.

Section 6.1 introduces the potential threat posed by filters to automatic face verification and deepfake detection networks. Due to the innovative nature of our research, no biometric datasets beautified via social media filters were available; therefore, we created our datasets, the process of which is explained in Section 6.2. Section 6.3 proposes to quantify the effect of each filter by using SSIM to define their *aggressivity*. In Section 6.4, we present the technologies evaluated in our experiments and the performance assessment of the face recognition and deepfake detector solutions. In addition, we conduct a subjective evaluation to assess the ability of a human observer to discern between deepfake and real videos before and after applying beautification filters. Finally, Section 6.5 provides a summary of the contributions made in this chapter.

The research question that this chapter aims to answer is: **RQ3**, i.e. *Do beautification filters pose a threat to the integrity of our existing biometric systems?*

6.1 Introduction

Creating, sharing and visualizing videos has become a daily activity for mobile users in the past decade. Social media platforms offer a diverse range of tools known as "filters" designed to automatically enhance a user's image,

requiring minimal or no user proficiency. Some types of filters, the so-called beautification filters are designed to tweak different facial features such as skin, lips, eyes, and nose to enhance the user's beauty. Among those, some filters cause more significant deformation of facial features compared to others since when contrasting original faces with beautified ones, some changes are more easily perceived than others. We refer to this characteristic as the *agressivity* of a filter.

Filters that alter facial biometric traits such as eyes and face contour, often subtly and imperceptibly to the human eye, inadvertently pose a threat to our face verification systems. A face image refashioned using these techniques could be taken as original or unprocessed for official documents, e.g. passport ID picture. Moreover, if telling the difference between a real video and a deepfake is difficult, the proliferation of beautification filters on social networks makes it nearly impossible to differentiate between a real video, a video enhanced by a filter, and a video with its original identity replaced.

6.2 FFMF and Celeb-DF-B Datasets

In this Section, we introduce the social media filters chosen for face beautification, the protocol employed in the creation of the FFMF [D3] and the Celeb-DF-B [D4] datasets and its final composition.

The filters employed in our dataset creation are chosen from the three most popular SNs[1]: Instagram, Snapchat, and TikTok. Instagram (Ig) is the social network that offers its users a wider variety of filters to apply to their multimedia content in real time. Instagram filters can be creator-sourced, opening the door to their users to create those effects for the general public use. Snapchat (Sc) was a pioneer social media in the use of filters beyond CAF. Although their app proposes a large range of facial filters, most of them are based on AR elements by adding extra objects to the subject, like hearts, flower crowns, or puppy ears. Ideally, in our study filters should modify the facial features slightly and/or blur the skin without noticeably modifying the facial traits. Tiktok (Tk) is a mobile app that focuses on the creation of short videos to which different effects can be applied such as music or beautification filters. The main constraint of their filter offer is the limited amount of FFMF that can be applied to existing multimedia content.

Besides being popular, the selected filters can be applied to already existing images or videos, an indispensable requirement for our study as the filtered images are created by modifying existing face images from biometric datasets.

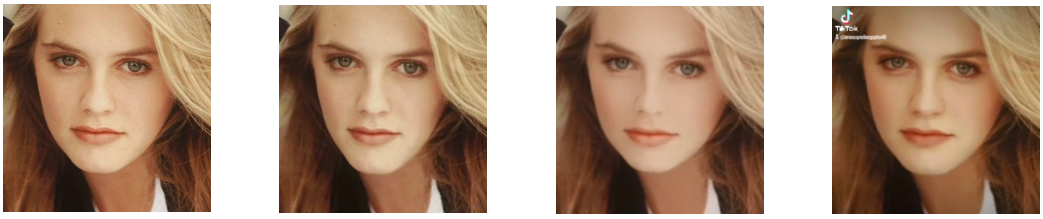
[1]Since 2012, Facebook and Instagram belong to the same group thus sharing features such as filters.

Table 6.1: Summary of the selected filters from Instagram, Snapchat, and TikTok. Includes type of filter according to the classification presented in Chapter 2 and face features modified by it. All filters except "Thinner_face" include Colour Adjustment.

SN	**Filter Name**	**F. Category**		**Modified Face Feature by F.**			
		SF	ARF	Contour	Eyes	Nose	Lips
Ig	Thinner_face			x	x		
	BROWN	x	x			x	
	California dreamin'	x				x	x
	Relax! You Pretty!	x				x	x
	Hawaii Grain	x	x		x	x	x
	Glam Grain	x	x		x	x	x
Sc	Fresh vibes	x	x	x	x	x	x
	Fresh light	x	x	x	x	x	x
	Mellow glow	x	x	x	x	x	x
Tk	Belle	x	x				
	Spring glow	x		x		x	

In Table 6.1 the selected SN filters are presented as well as various types of filter modifications and different biometric features that can be altered by the FFMF. Trait modifications were assessed by visual inspection of pixel differences between original and filtered images. Examples of resulting images, once beautification filters are applied, are shown in Fig. 6.1. Unlike the other SNs, TikTok automatically adds a visible watermark image to all processed images/videos indicating the app logo and account identity (see Fig. 6.1). The watermark's position varies across all images in the TikTok FFMF dataset, consistently positioned in the corners of the images, thus never obstructing the central facial regions. By employing the Viola-Jones algorithm to detect and crop all faces in the expanded FFMF dataset, we were able to eliminate the TikTok watermark, which could have impacted the evaluation process in the next Section.

In Figure 6.2 we present the pipeline followed for the "beautified" images creation. The upload of content and application of filters on social media platforms has some requirements: filters have to be applied online, manually, and to one image or video at a time, there is a maximum number of seconds (s) and a restriction of 30 frames per second (fps) for an uploaded video. Following all these prerequisites and in order to ease the process of manually applying the filters, we created 15s videos with the images, with each still image displayed for 0,3 seconds (i.e. 10 frames). The videos are then passed



Original image | Thinner_face | Mellow Glow | Belle

Figure 6.1: Example of original and beautified images, one filter for each SN: Thinner_face (Ig), Mellow Glow (Sc) and Belle (Tk).

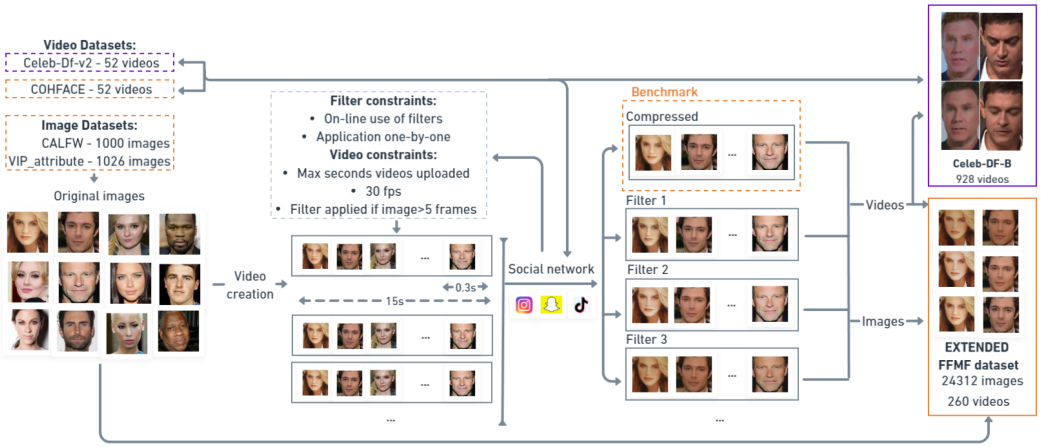


Figure 6.2: Overview of the creation of new data for the FFMF and Celeb-DF-B datasets. The images from the source datasets are arranged in videos and passed through different social media where all the original content is processed and filtered. The videos are passed directly to the desired social network.

through the different social networks to be filtered. After that, one frame per still image is selected. The videos from the source datasets are directly uploaded to the SN and beautified with Instagram filters.

FFMF Dataset

The **FFMF dataset [D3]** is composed of processed images and videos initially presented in the three publicly available datasets CALFW [2], VIP_attribute [39] and COHFACE [49]. The VIP_attribute and COHFACE datasets were presented in Chapter 3. The Cross-Age LFW (CALFW) [143] is an improved

[2]http://whdeng.cn/CALFW/?reload=true#download

version of the LFW face dataset [145]. It incorporates additional face pairs with age disparities, thereby introducing age diversity and intra-class variance while maintaining the same identities as the original LFW dataset. CALFW is a public benchmark for face verification with face photographs designed for studying the problem of unconstrained FR. Comprising 4,025 individuals, each person in the CALFW dataset is associated with up to four images. For our dataset creation, all the 1026 images presented in VIP_attribute, a selection of 52 face videos from the COHFACE dataset and a gender-balanced subset composed of 1000 images from the CALFW were selected. According to their source dataset, the images and videos of the extended FFMF are referred to as 1) FFMF-VIP, 2) FFMF-CALFW and 3) FFMF-COHFACE.

Three different categories of visuals are considered in the FFMF dataset, original, uploaded, and beautified (using various filters from the different social networks). "Original" visuals have not undergone any processing. "Uploaded" visuals were only uploaded (without any filters applied) to the three social networks, although the platform may perform some processing to resize the images and videos. When content is uploaded to SNs, it goes through operations such as image compression, resizing, and cropping, which has been proved to have a negative impact on facial processing tasks such as face recognition [146]. Therefore, the original visuals were uploaded to and downloaded from the different social media platforms obtaining the "uploaded" image category in the dataset, which is used in Section 6.4.3 and Chapter 7 as a baseline. Finally, "beautified" visuals are those to which filters have been applied.

The FFMF dataset is composed of 24312 images and 260 videos belonging to the three categories, original, uploaded, and beautified. All images present in FFMF-VIP and FFMF-CALFW are "uploaded" to the three social networks and "beautified" via "Thinner_face", "Relax! You Pretty!", "Glam Grain", "Fresh vibes", "Fresh light", "Mellow glow", "Belle" and "Spring glow". The videos in FFMF-COHFACE have been "uploaded" to Instagram and "beautified" by applying the filters "Thinner_face", "Relax! You Pretty!" and "Hawaii Grain".

Celeb-DF-B Dataset

The **Celeb-DF-B [D4]** dataset is composed of a subset of videos initially presented in the Celeb-DF [147] dataset and their self-beautified counterpart. The Celeb-DF dataset consists of 590 real and 5639 DeepFake videos with an average duration of 13 seconds per video and a standard frame rate of 30 frames per second. The real videos are sourced from publicly accessible YouTube content corresponding to interviews featuring 59 celebrities. Among

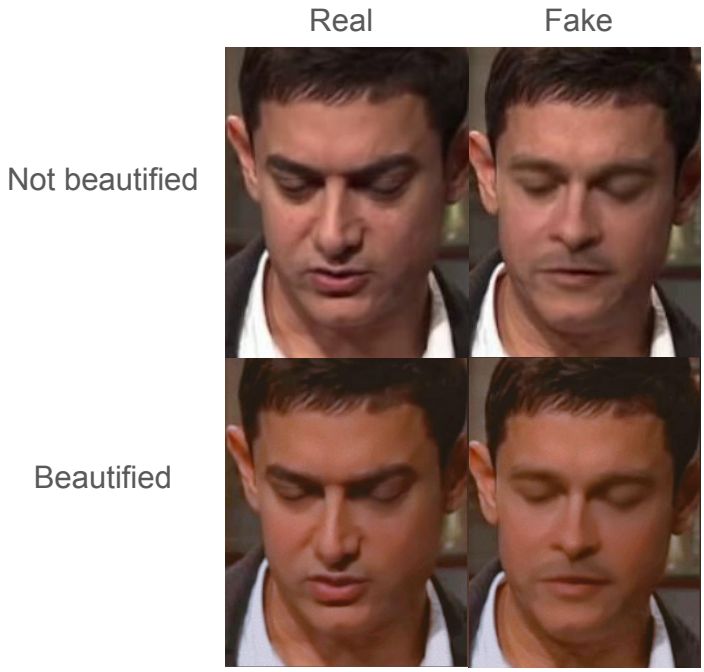


Figure 6.3: Frames extracted from four distinct videos within the Celeb-DF-B dataset are depicted here. The top row showcases frames from non-beautified videos, whereas the bottom row exhibits beautified frames. The frames in the left and right columns pertain to real and fake videos respectively.

these, for the creation of the Celeb-DF-B dataset, we chose a subset consisting of 232 real and 232 fake videos. The selection of videos followed three criteria: 1) an equal sampling from each identity in the real videos; 2) pairing each real video with a fake counterpart created through FaceSwap; and 3) maintaining a balance between the source and driving identities of the selected fake videos.

Each of the 464 non-beautified selected videos is beautified with one of the four following filters: "BROWN", "California dreamin'", "Relax! You Pretty!" and "Hawaii Grain" resulting in the creation of 928 videos that constitute the Celeb-DF-B dataset. Example frames of the videos belonging to Celeb-DF-B dataset are displayed in Figure 6.3.

6.3 Filter Aggressivity Assessment

In this Section, we propose the Structural Similarity Index (SSIM), an index commonly used for assessing the perceived changes between two compared images, as a metric to quantify the *aggressivity* of a filter. We define a filter as

Table 6.2: SSIM coefficient between the original and processed images of the FFMF extended dataset. Results are apportioned by gender: female (F) and male (M).

SN	Img. Processing	CALFW (F)	CALFW (M)	VIP (F)	VIP (M)
Ig	Uploaded	97	97	98	98
	Thinner_face	92	92	91	92
	Relax! You Pretty!	80	82	81	81
	Glam Grain	77	78	77	77
Sc	Uploaded	96	96	96	96
	Fresh vibes	82	82	81	82
	Fresh light	91	91	91	92
	Mellow glow	89	89	89	89
Tk	Uploaded	97	97	98	98
	Belle	94	94	93	93
	Spring glow	94	94	93	94

more *aggressive* than another if it modifies a larger number of facial features than others. The SSIM index is based on the combination of 3 different terms: luminance $l(I, J)$, contrast $c(I, J)$, and pixel structure $s(I, J)$. Given two RGB images I and J:

$$SSIM(I, J) = [l(I, J)]^{\alpha}\dot{[}c(I, J)]^{\beta}\dot{[}s(I, J)]^{\gamma} \tag{6.1}$$

$$l(I, J) = \frac{2\mu_I\mu_J + C_1}{\mu_I^2 + \mu_J^2 + C_1} \tag{6.2}$$

$$c(I, J) = \frac{2\sigma I\sigma_J + C_2}{\sigma_I^2 + \sigma_J^2 + C_2} \tag{6.3}$$

$$s(I, J) = \frac{\sigma_{IJ} + C_3}{\sigma_I + \sigma_J + C_3} \tag{6.4}$$

with μ_I, μ_J, σ_I, σ_J and σ_{IJ} the mean of I and J, the standard deviation of I and J, and the covariance of I and J, respectively. The luminance index compares the average brightness of the pixels in the reference and distorted images. The contrast index and the structural index, juxtapose the standard deviation and the covariance of the pixel values in the reference and distorted images, respectively.

We present the SSIM index between the original and processed ("uploaded" and "beautified") images on the extended FFMF dataset in Table 6.2. For the three SNs, the images already present a degradation when uploaded to the

platforms without any beautification filter application, with Snapchat images having a slightly higher penalization. We can observe that the TikTok filters (Belle and Spring glow) and the Instagram filter Thinner_face have a higher similarity with the original image when compared to other filters. This is consistent with the information presented in Table 6.1, where we can see how those filters modify a smaller number of biometric features. The difference is greater for the Snapchat filters Fresh light and Mellow glow, which modify a bigger number of facial traits. Nevertheless, Fresh vibes strongly penalizes the SSIM score indicating that although a small number of features is modified, their variations are significant. Finally, a strong impact of Instagram filters (Relax! You Pretty and Glam Grain) on facial traits is observed (lowest SSIM score). No significant SSIM discrepancies in terms of gender are observed, meaning that filters equally impact the SSIM of the female and male subjects.

6.4 Impact in Face Biometrics

In this section, we present the technologies evaluated in our experiments, the evaluation protocol followed, the implementation details, and the metrics used in our study. Additionally, we discuss the experimental results of the assessment conducted on the datasets presented in Section 6.2.

6.4.1 Methodology

Five AI-based networks were selected for our experiments: two state-of-the-art face recognition systems (ArcFace and MagFace) and three high-performance deepfake detectors(CADDM, RECCE, and FTCN).

Face Verification

ArcFace [83], is a face recognition solution based on a ResNet architecture. ArcFace, or Additive Angular Margin Loss, is a loss function designed to explicitly optimize the feature embedding. It aims to enforce higher similarity among intra-class samples while maintaining diversity among inter-class samples. This is particularly important for addressing the performance gap in deep face recognition caused by significant intra-class appearance variations. ArcFace has demonstrated high verification performance on well-known face recognition benchmarks, namely LFW [145], establishing itself as a state-of-the-art technique for FR.

MagFace [148], a revisited version of ArcFace designed to address the problem of large intra-class variability in subjects' faces, which is stronger in

unconstrained acquisition scenarios. MagFace integrates quality measures into the face representation through the magnitude-aware angular margin loss. By simultaneously enforcing the direction and magnitude of cosine distance, the learned face representation becomes more robust to the variability observed in real-world faces.

Deepfake Detection

CADDM [149] detects traces of forgery at the frame level. Initially, the image is passed through an EfficientNet-b4 [150] backbone to extract useful features for the classification task. Subsequently, it identifies forgery locations across various scales via an artifact detection module trained with a custom Multi-scale Face Swap algorithm to generate forgery location ground truth. The classification score of the video is computed by averaging the scores obtained from individual frames. The purpose of this architecture is to prioritize the detection of local forgeries rather than solely learning facial distributions enhancing the performances when detecting fakes of previously unseen faces.

RECCE [151] is an encoder-decoder-based model. The encoder is based on Xception [152]. The reconstruction network has been trained in an unsupervised manner to learn the representation of real faces. In this model, face frames are passed through the encoder-decoder architecture. Then, encoder and decoder features are combined together with the residual images (i.e. the difference between the reconstructed and the original frame) to classify each frame as fake or genuine. The classification score for the video is computed as the average score across all frames.

FTCN [153] is a model trained to detect temporal inconsistencies in videos. As deepfakes are generated frame by frame, they often exhibit temporal inconsistencies. The FTCN network utilizes a ResNet50 3DCNN backbone to extract temporal features and employs Temporal Transformers as a classifier. Therefore, FTCN does not analyze each image independently but instead focuses on the sequence of frames, enabling it to identify temporal inconsistencies characteristic of deepfake videos.

Subjective Evaluation: We performed a subjective evaluation of deepfake videos, using a web-based framework for crowdsourcing experiments. The primary objective of this subjective test was to investigate whether the utilization of such filters presents challenges for human observers when distinguishing between the authenticity of deepfake videos and real videos.

6.4.2 Experimental Setup

Metrics

To assess face verification, the metrics and terminology recommended by the ISO/IEC standard [154] are adopted: False match rate (FMR): proportion of zero-effort impostor attempt samples falsely declared to match the compared non-self template; False non-match rate (FNMR): proportion of genuine attempt samples falsely declared not to match the template of the same characteristic from the same user supplying the sample; Detection error trade-off (DET) curve; modified Receiver Operating Characteristic (ROC) curve which plots error rates on both axes (false positives on the x-axis and false negatives on the y-axis). In addition, we present accuracy and Area under the Curve (AUC), metrics that are also often reported in works on FR.

To evaluate the three selected deepfake detectors, we analogously compute the video-level AUC of the ROC curve and the False Negative Rate (FNR), i.e., the proportion of fake videos recognized as genuine, which, in a real-case scenario, is desirable to minimize. Additionally, we analyze the histogram of the classification scores before and after beautification to gain a better understanding of the behavior of deepfake detectors on beautified videos. In the subjective evaluation, we report accuracy and recall. Recall, also known as sensitivity or true positive rate, measures the ability of a classifier to correctly identify positive instances among all actual positive instances. In the context of deepfake detection, a higher recall implies that the deepfake detection model or human evaluators are better at spotting deepfakes when the videos are beautified.

Evaluation Protocol

To assess the impact of digital beautification on face verification, the models are challenged in our FFMF dataset presented in Section 6.2. The following experimental protocol was designed: each face analysis model is evaluated (i) on the "original" images from FFMF; (ii) on the "uploaded" images (without beautification filters); (iii) on the "beautified" images. The results obtained from (i), (ii), and (iii) are then compared to assess the impact of the different filters. In particular, experiment (ii) is designed to isolate the impact of uploading images to social networks (SNs) – which often undergo image compression, resizing, and cropping operations – from the actual impact of beautification.

We test the performance of each deepfake detector in the Celeb-DF-B presented in Section 6.2, following the evaluation process defined by Dong *et al.* [149]. We extract 32 frames at equal intervals to obtain 32 classification

scores. Each evaluation score represents a real number between 0 and 1 for real and fake videos, respectively. The video score is then computed as the average of all the individual scores. FTCN, on the other hand, extracts a sequence of N consecutive frames from the video. To maintain consistency with the evaluation of CADDM and RECCE, we set $N = 32$. In our study, we define the positive class as 'fake videos' and the negative class as 'genuine videos'.

For the subjective test, a total of 112 videos (56 real and 56 deepfakes) were selected from the Celeb-DF-B dataset. For fake videos, 7 videos were randomly chosen for each type of beautification filter, resulting in 28 videos. The same videos were included without the filter in the subjective test dataset. For real videos, 7 videos were selected for each filter type, and these videos were also included without filters in the subjective test dataset. The test protocol involved presenting human evaluators with cropped face regions and extending the boundary by an additional 100 pixels into the background. Before the evaluation, participants received comprehensive explanations of the test procedures and completed practice tests to ensure their understanding. To optimize efficiency and prevent fatigue during the evaluation, we divided the test dataset randomly into three batches. This approach allowed participants to complete each test in separate sessions, with breaks in between. On average, each test batch lasted approximately 15 minutes, consistent with the standard recommendations [155].

Implementation details

Our ArcFace implementation is based on the one provided by the InsightFace project[3]. The network was pre-trained on the MS1MV2 [156] dataset. We studied different setups during our experiments, such as different backbones (ResNet50 vs ResNet100). Our MagFace implementation with iResNet100 (improved residual network) architecture as backbone is based on the GitHub project by Irving Meng[4]. The model is pre-trained on the MS1MV2 dataset [156].

A largely adopted state-of-the-art protocol for Face Verification is used [145]. Ten folds of images are created by randomly selecting images from the dataset. Each fold contains 300 matching pairs and 300 non-matching pairs, for a total of 6000 face comparisons (10×600). This random selection is uniform across datasets of the "original", "uploaded" and "beautified" facial images. Face landmarks are computed for each face image using RetinaFace [157] face detector to take into account the possible displacement of facial features

[3]https://github.com/deepinsight/insightface
[4]https://github.com/IrvingMeng/MagFace

Table 6.3: Type of data from the FF++ dataset used in the training of each deepfake detector considered.

Model	Compression	Seen Face Manipulation	Seen fake samples
CADDM [149]	Raw	DF, F2F, FSh, FS, NT	Yes
RECCE [151]	c23	DF, F2F, FSh, FS, NT	No
FTCN [153]	c23	DF, F2F, FS, NT	Yes

due to the application of the filters. In the event that the face landmark estimation via RetinaFace fails for one of the face images in the list of pairs, the corresponding comparison is excluded from the test.

Our implementations of CADDM[5] RECCE[6] and FTCN[7] are based on publicly available GitHub projects. All three deepfake detector models use a backbone trained on ImageNet to extract features and are trained on FaceForensics++ [158] (FF++) for the deepfake detection task. FF++ contains respectively 5000 and 1000 fake and real videos divided into three subsets: train, val, and test. Five manipulation techniques were used to generate the fake videos. They are either face reenactment (Face2Face: F2F, NeuralTexture: NT) or FaceSwap (Deepfake: DF, FaceSwap: FS, FaceShifter: FSh) based methods. All the 6000 videos exist in 3 versions: *raw*, *High-Quality (c23)*, and *Low Quality (c40)*. Table 6.3 gives a summary of the specific data seen by each model during their training on FF++. For more information about the training of the three models, please refer to their corresponding publications.

The subjective evaluation involved 21 participants with diverse backgrounds. Each video was shown to the participants three times consecutively. After viewing each video, following a procedure similar to that of Korshunov et al. [159], participants were asked, "Is the person's face in the video real or fake?" They were then asked to identify the specific features or characteristics that influenced their judgment regarding the video's authenticity. The available feature options included: 1. Face contour, 2. Shadow inconsistency, 3. Inconsistency between eyes, 4. Eye blinking, 5. Mouth, 6. Teeth, 7. Lip motion, 8. Head motion, 9. Face/body mismatch, 10. Contextual mismatch, 11. Skin texture, and 12. Video quality.

[5]https://github.com/megvii-research/CADDM
[6]https://github.com/VISION-SJTU/RECCE
[7]https://github.com/yinglinzheng/FTCN

Table 6.4: Assessment of the impact of beautification filters on *face verification* on the FFMF dataset. Accuracy, corresponding standard deviation (±) and the Area Under the ROC Curve (AUC) are reported. The *higher* the value, the better.

		FFMF-CALFW		**FFMF-CALFW**		**FFMF-CALFW**	
		ArcFace ResNet 50		**ArcFace ResNet 100**		**MagFace iResNet 100**	
Experiment		**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**
Original		0.9351	0.9030±0.0041	0.9444	0.9074±0.0059	0.9810	0.9417±0.0076
Ig	Uploaded	0.9308	0.9047±0.0086	0.9437	0.9123±0.0071	0.9805	**0.9418±0.0063**
	Thinner_Face	0.9381	0.9090±0.0053	0.9481	0.9157±0.0056	0.9802	0.9373±0.0070
	Relax! You Pretty!	0.9379	0.9030±0.0087	0.9464	0.9145±0.0089	0.9769	0.9245±0.0099
	Glam Grain	0.9322	0.8963±0.0054	0.9424	0.9096±0.0073	0.9773	0.9295±0.0095
Sc	Uploaded	0.9398	0.9085±0.0048	0.9410	0.9203±0.0056	**0.9818**	0.9398±0.0060
	Fresh light	0.9401	0.9118±0.0092	0.9543	**0.9231±0.0083**	0.9783	0.9300±0.0090
	Fresh vibes	0.9351	0.9046±0.0113	0.9488	0.9139±0.0105	0.9768	0.9315±0.0078
	Mellow glow	0.9300	0.9103±0.0100	0.9441	0.9195±0.0087	0.9753	0.9238±0.0087
Tk	Uploaded	0.9340	0.9045±0.0063	0.9478	0.9149±0.0064	0.9814	0.9398±0.0081
	Belle	**0.9493**	**0.9160±0.0074**	**0.9584**	0.9216±0.0079	0.9793	0.9360±0.0084
	Spring glow	0.9418	0.9038±0.0093	0.9498	0.9159±0.0087	0.9771	0.9317±0.0091

Table 6.5: Assessment of the impact of beautification filters on *face verification* on the FFMF dataset. Performances are reported in terms of FNMR (%) at fixed values of FMR: 10 - 100 - 1000. The *lower* the value, the better.

		FFMF-CALFW			**FFMF-CALFW**			**FFMF-CALFW**		
		ArcFace ResNet 50			**ArcFace ResNet 100**			**MagFace iResNet 100**		
Experiment		**10**	**100**	**1000**	**10**	**100**	**1000**	**10**	**100**	**1000**
Original		13.35	18.95	23.61	12.38	17.63	21.98	4.57	15.90	40.60
Ig	Uploaded	13.11	18.33	24.4	12.2	17.07	24.5	4.70	16.07	34.20
	Thinner_Face	11.84	18.17	24.7	10.89	15.57	**30.05**	4.67	18.00	43.40
	Relax! You Pretty!	12.97	18.68	22.95	**12.48**	16.19	25.37	5.87	5.47	34.47
	Glam Grain	13.88	**21.58**	**32.47**	12.34	**17.77**	23.48	5.40	20.17	40.67
Sc	Uploaded	11.98	17.92	23.07	11.14	15.2	22.27	4.57	15.63	34.13
	Fresh light	12.17	17.03	19.83	10.84	14.51	18.6	5.20	19.73	2.23
	Fresh vibes	13.59	18.45	24.05	11.87	16.21	21.01	5.50	18.13	45.70
	Mellow glow	13.08	18.24	26.73	11.16	15.43	21.29	**6.00**	16.63	34.37
Tk	Uploaded	**14.16**	18.78	22.04	11.35	16.43	19.94	.43	15.87	41.33
	Belle	0.71	6.47	8.32	.53	4.3	7.72	4.93	17.60	44.50
	Spring glow	11.4	18.54	22.98	11.51	16.19	20.98	5.23	**20.70**	**46.67**

6.4.3 Results

Face Verification

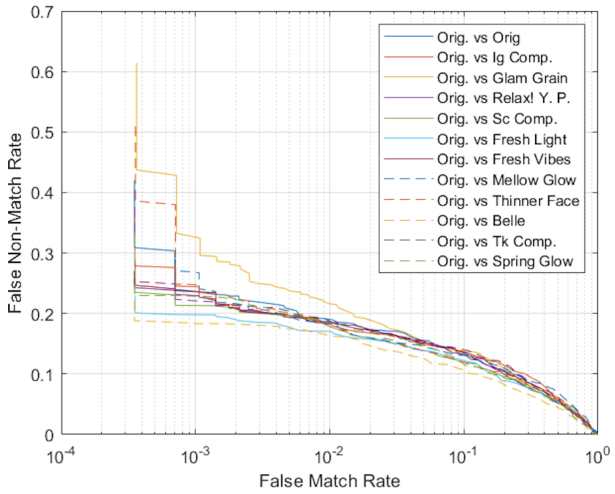
Table 6.4 reports the performance in terms of accuracy and AUC. The highest values are indicated in **bold** text. Mellow glow, Glam Grain, "uploaded" by Sc, and "original" all achieve the lowest accuracy or AUC values. Interestingly, for

ArcFace w/ ResNet100, the lowest performance in terms of accuracy is that of the original images, which means that somehow the application of the filters has improved the accuracy of face recognition. In line with this statement, if we look at the *positive* impact, and thus the higher accuracy and AUC values, it is the Belle filter that achieves the best results overall, followed by Fresh light, "uploaded" by SC, and "uploaded" by Ig. This corroborates what is reported in Ueda *et al.*[80] about light make-up, namely slight facial modifications aimed at beautification can even improve recognition performance.

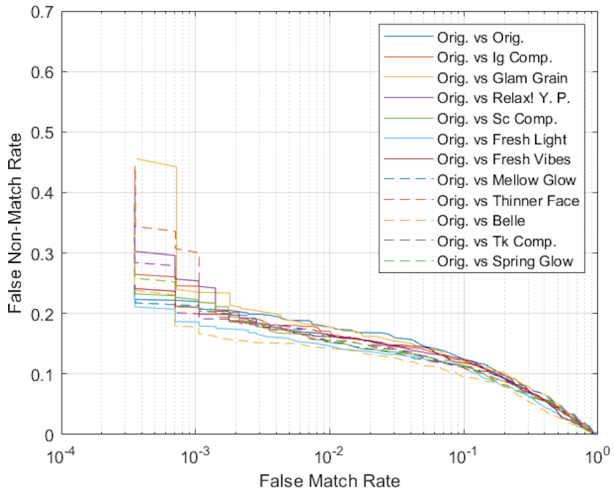
The results presented in Table 6.5 and illustrated in Fig.6.4 report the performances of face verification in terms of FMR and FNMR. They should be read as the rate of authorized users wrongly rejected (FNMR) at fixed rates of impostors (i.e. unauthorized users) wrongly accepted (FMR). It can be observed that at FMR1000, the difference between the filters is wider. This means that, for example, for a system with higher security, i.e. that allows only 1/1000 impostor to access the system, the impact of filters is more significant. Regarding ArcFace, the results, both for negative and positive impact, converge most clearly on two filters. Concerning the negative impact, and thus the highest error rates, Glam Grain, which is the most aggressive filter in our selection, achieves the overall highest error rates, with Thinner_Face, Relax! You Pretty!, and "uploaded" by Tiktok achieved the highest error rates for some of the tests too. The largest performance drop is obtained at FMR1000 with Glam Grain for ResNet50 and Thinner_Face for ResNet100, where about 10% more authorized persons would be rejected when using the Glam Grain/Thinner_Face filter compared to the original. As for the positive impact, the filter Belle, which is the least aggressive, clearly improves the performance of face verification for all tests. This means that images processed with this filter are less likely to be falsely rejected in line with Ueda *et al.*[80] and Table 6.4. As for MagFace, the largest drops in performance are achieved by Mellow glow, and Spring glow. While the largest improvements are obtained by Relax! You Pretty!, Fresh light, and "uploaded" by Tk. This indicates that the impact of compression and beautification filters depends on the architecture and the loss function used by the FR model. Belle, the filter that shows a higher SSIM value, has been proven to act as light makeup thus increasing the user's feminity leading to an increase in face recognition performances, once more in line with the findings of Ueda *et al.* [80].

Deepfake Detection

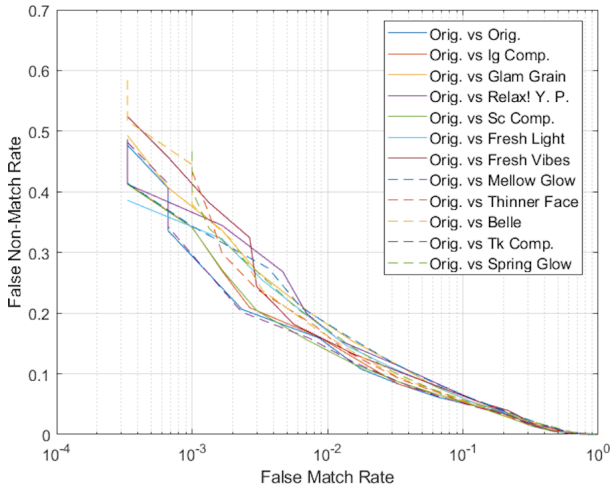
In Table 6.6 and Figure 6.5, we present various results of our experiments. In Table 6.6, we observe that all detectors suffer a drop of approximately



(a) ArcFace w/ ResNet50



(b) ArcFace w/ ResNet100



(c) MagFace w/ iResNet100

Figure 6.4: *Face Verification* assessment on the FFMF dataset. Detection error tradeoff (DET) curve for (a) ArcFace w/ ResNet50; (b) ArcFace w/ ResNet100; and (c) MagFace w/ iResNet100.

Table 6.6: Assessment of the impact of beautification filters on *deepfake detection* on the Celeb-DF-B dataset. AUC score of each detector w/o and w/ beautification. The *higher* the value, the better.

Model	AUC	
	w/o beautification	**w/ beautification**
CADDM [149]	0.91	0.76 (↓ 0.15)
RECCE [151]	0.81	0.66 (↓ 0.15)
FTCN [153]	0.80	0.64 (↓ 0.16)

15% in AUC when tested with beautification filters. In Figure 6.5 (a), we see how depending on the detector used the beautification process may increase the FNR. Specifically, we observe that beautified videos reduce the FNR for CADDM and FTCN. However, for RECCE, the False Negative Rate is higher for beautified videos. This presents a significant issue, as fake videos may appear authentic due to the simple application of a beautification filter. In real-world scenarios, minimizing the FNR, i.e., the proportion of fake videos detected as real, is crucial. In contrast to CADDM and FTCN, RECCE did not encounter any fake videos during its training as presented in Table 6.3, and thus, it did not learn any specific features associated with face manipulation. Consequently, even if beautification introduces minor artifacts, it removes some of the manipulation introduced by deepfakes. However, supervised trained models such as CADDM and FTCN can detect these artifacts.

To better understand the behavior of deepfake detectors on beautified videos, we analyzed the histogram of the classification scores before and after beautification. In Figure 6.6, we illustrate the difference in the distribution of the classification scores of the deepfake detectors on Celeb-DF-B for beautified and non-beautified videos. For CADDM and FTCN, we can observe higher confidence scores for the fake label of the beautified videos, indicating they are more likely to be detected as fake. On average, all the confidence scores of the videos are shifted by +0.1 and +0.3, respectively, for CADDM and FTCN after beautification. However, the behavior is slightly different for RECCE. On average, beautified videos appear more authentic than the original ones, with an average score shift of -0.07. This trend is illustrated in Figure 6.6. In summary, for RECCE, beautified videos tend to appear more real than their non-beautified counterparts.

Subjective Evaluation of Deepfakes

Table 6.7 presents the results of the subjective assessment. The results indicate that human accuracy for non-beautified videos is higher (69%) than

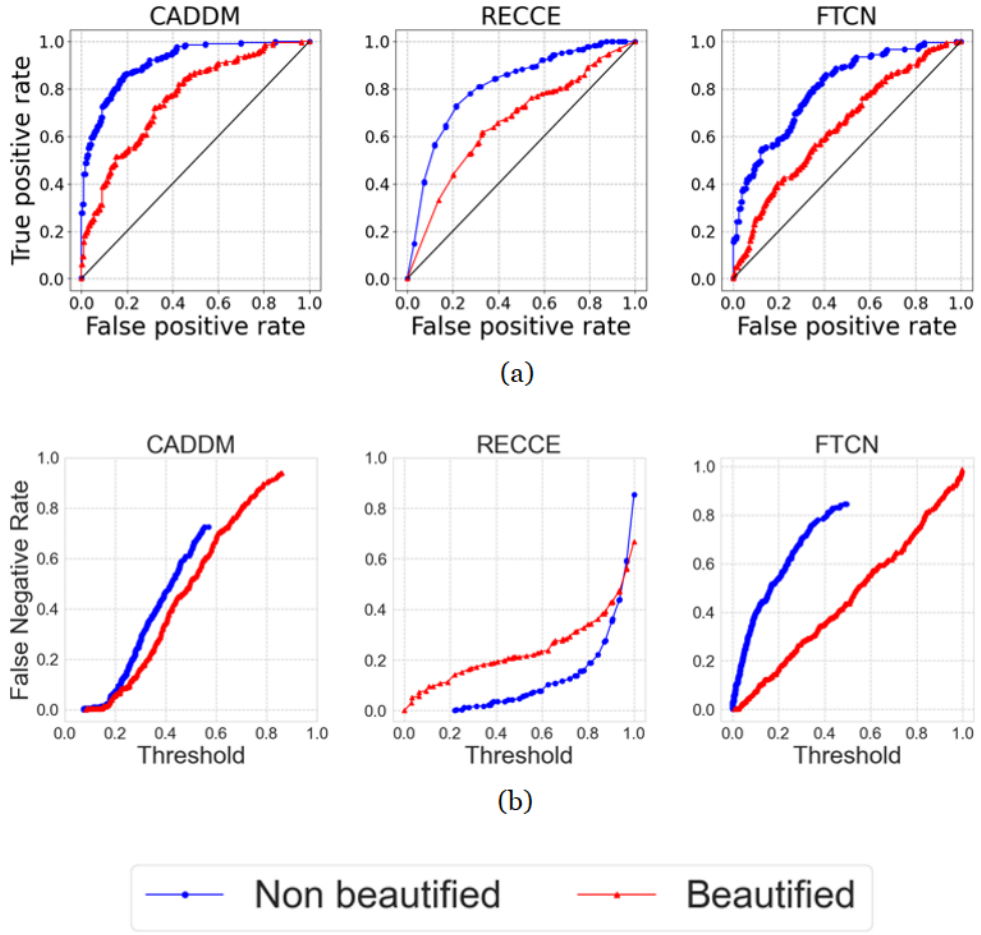


Figure 6.5: Assessment of the *deepfake detectors* on the Celeb-DB-F dataset. Each group of images presents (a) the video-level AUC of the ROC curve, and (b) the FNR for different classification score thresholds.

Table 6.7: Subjective evaluation results of the impact of beautification filters on deepfake detection on the Celeb-DF-B dataset for beautified and non-beautified videos. The *higher* the value, the better.

Metric	**Non-beautified**	**Beautified**
Accuracy	0.69	0.66
Recall	0.70	0.76

for beautified videos (66%), suggesting that human judgments are more effective at distinguishing between real and fake videos when no beautification is applied. Furthermore, we can observe a significant contrast in recall rates

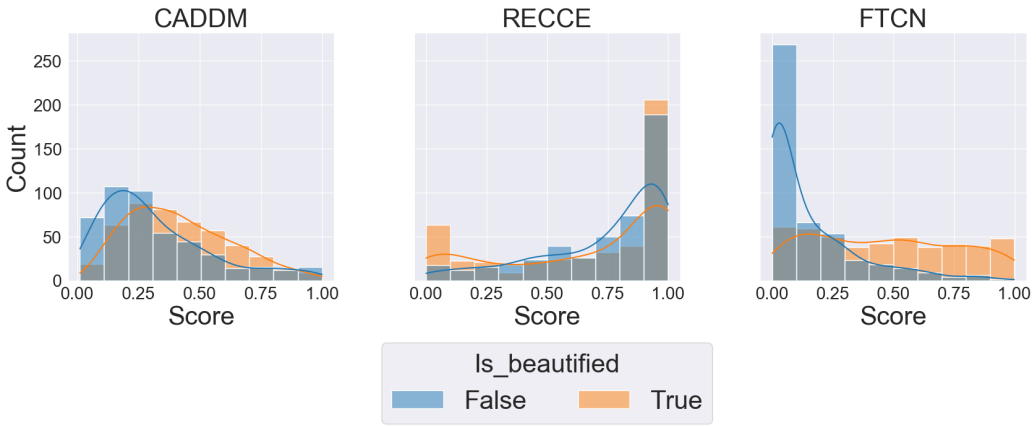


Figure 6.6: Histogram of the classification scores of the *deepfake detectors* on the Celeb-DB-F dataset.

between beautified and non-beautified videos. The recall rate for beautified videos is 76%, while for non-beautified videos, it is 70%. The increased recall rate in our study implies that evaluators perform well at identifying deepfakes when beautification filters are present. However, the observed accuracy rates suggest that while human evaluators improve in detecting deepfakes with applied filters, they also tend to misclassify more genuine videos as fake in this scenario. This highlights the impact of beautification filters on human detection capabilities: they not only aid in recognizing deepfakes but may also lead to a higher rate of false positives, where non-deepfake videos are mistakenly identified as unreal.

In the subjective evaluation, participants were also tasked with identifying the specific features or characteristics that played a role in their decision regarding the video's authenticity. Among the provided feature options, the inconsistency between eyes stood out as the most frequently noted feature in both beautified and non-beautified videos. An interesting finding is that many participants highlighted alterations in skin texture as a factor influencing their categorization of videos as fake, with a higher percentage observed in beautified videos, as presented in Figure 6.7.

6.5 Summary

In this chapter, we investigate the effects of beautification filters in face biometrics by conducting experiments on FFMF and Celeb-DF-B, two novel datasets created by applying popular social media beautification filters to a subset of images and videos from publicly available datasets. We provide

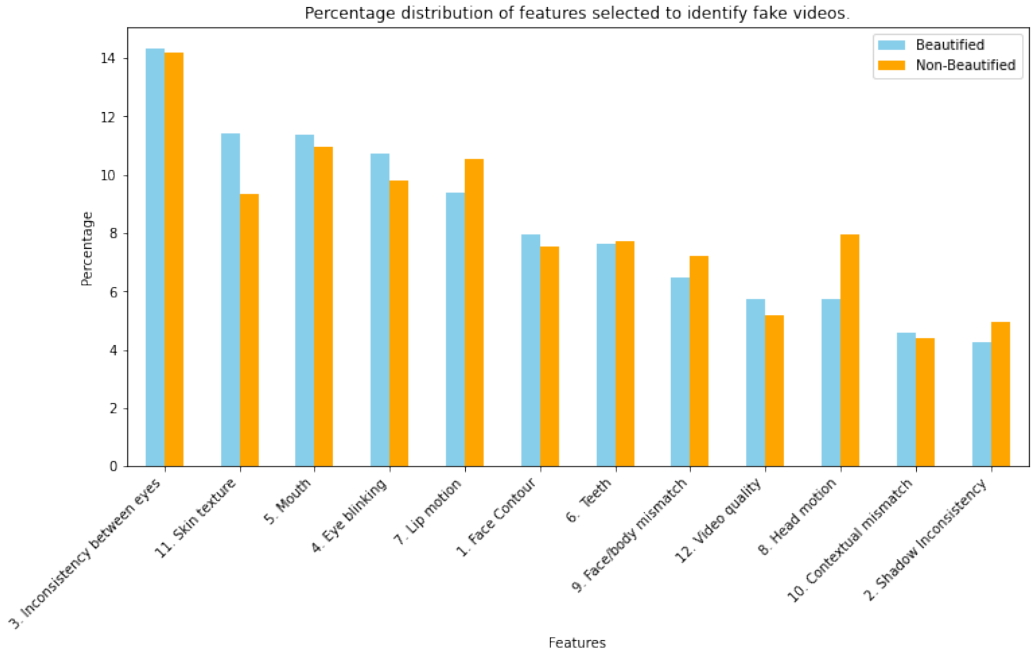


Figure 6.7: Prominent features identified by the users in the subjective test evaluation as influencing to their decisions.

details on the protocol used for creating these datasets and introduce cosine similarity as a metric to objectively measure the aggressiveness of each filter on face images. We conducted an assessment using two state-of-the-art face recognition systems and three state-of-the-art deepfake detectors. The results indicate that these filters significantly alter the behavior of the studied networks. In the context of human-level performance, we proved that the use of filters similarly influences human decision-making, affecting the accurate categorization of videos as either real or fake.

In the next section, we will assess the performance of soft and hidden biometric estimators when confronted with the variability introduced by filters in the datasets.

Chapter 7

Impact of Beautification Filters in Facial Soft Biometric Estimation

This Chapter extends the study of the impact of beautification via social media filters to other facial processing tasks beyond identity recognition. We evaluate the impact of beautification filters on soft and hidden biometric estimators, focusing on attributes such as gender, age, weight, and heart rate.

In Section 7.1, we provide the rationale for assessing other facial biometric networks in addition to traditional face recognition and deepfake detection tasks. Section 7.2 elaborates on the selected networks, the experimental setup, and the results of our study. Finally, in Section 7.3, the chapter concludes with an overview of the impact of filters on the various studied tasks.

The research question that this chapter aims to answer is: **RQ3**, i.e. *Do beautification filters pose a threat to the integrity of our existing biometric systems?*

7.1 Introduction

As we saw in Chapter 6, the application of filters to facial multimedia is a user-friendly practice, as it does not demand any prior expertise, unlike other image editing techniques. This accessibility makes filters highly approachable for the average social media user. This not only challenges the robustness of current identity verification methods but also raises important questions about the reliability of other facial processing models that can be affected in real-world scenarios, where such filters are commonly used.

Monitoring physiological parameters, such as weight or heart rate, is of

great importance to address an individual's health status and is beneficial not only for patients in critical situations but also for high-risk patients in home care and outdoor areas [160]. Gender, age, weight, and heart rate are routinely assessed in clinics and hospital emergencies as a first step in patient diagnosis. With the rise in popularity of telehealth, especially during pandemic times, those parameters can be monitored by a health professional remotely and thus its estimation can be compromised by the use of beautification filters.

7.2 Impact in Soft Biometrics

In this section, we provide a comprehensive overview of the technologies evaluated, including their implementation details. We also outline the metrics and evaluation protocol employed in our experiments, followed by a discussion of the results obtained.

7.2.1 Methodology

Two popular open-source gender estimators *DeepFace* and *cvlib* are adopted. *DeepFace* provides the most popular pre-trained models for face detection and FR along with its own models for gender classification. *cvlib* is a high-level open-source computer vision library for Python. It includes an AlexNet model trained for gender classification. Both *DeepFace* and *cvlib* return the labels "man", "woman", and associated probabilities, once a human face is passed to the gender models. As in Chapter 5, we adopt DEX [23] as a model for apparent age estimation. DEX is based on the VGG-16 architecture pre-trained on ImageNet. It extracts predictions from an ensemble of 20 age estimator networks from the subject's cropped face without explicitly using facial landmarks. For assessing the impact of FFMF on weight estimation, we selected likewise to the network presented in Chapter 3 a ResNet architecture with 50 layers and a final regression layer. Similarly, we assessed the effect of beautification filters on remote HR estimation through the network proposed in Chapter 3, a 3-dimensional CNN that directly estimates the HR from RGB face videos.

7.2.2 Experimental Setup

Metrics

Global accuracy and accuracy per class are presented in the gender classifier assessment. For age and weight, we provide the Me and StD of the difference between the computed age from the original and processed images in years

and the difference between the real and predicted weight in kg. We also report the MAE in years and kg respectively and the correlation coefficient ρ. Additionally, we include the Percentage of Acceptable Predictions for the weight estimation network. To assess the beautification filters' impact on HR estimation we employ similarly the StD in bpm of the heart rate error, the MAE in bpm and the Pearson's correlation coefficient.

Evaluation Protocol

Similar to the experiments presented in Chapter 6 for face verification, we assess the impact of digital beautification on facial processing tasks on the FFMF dataset and the following experiments are designed: the different facial processing tasks are evaluated (i) on the "original" images; (ii) on the "uploaded" images (no beautification filters); (iii) on the selected "beautified" images.

Implementation Details

For the gender classifiers, *DeepFace*[1] and *cvlib*[2], we use their official Github implementations.

Since the datasets CALFW and VIP_attribute are not annotated by age, we assess the impact of beautification filters on apparent age. The authors of the CALFW dataset ranked the images of each subject from younger to older [143] and they estimated the age of each image using DEX. In our experiments with DEX, the implementation provided in GitHub by *siriusdemon* is used[3].

For assessing the impact of filters on facial-based weight estimation, the experiments are performed just on the FFMF-VIP dataset. Following the same protocol as presented in Section 3.2.2, 800 original images (400 female and 400 male) are selected for the training. The results of our experiments are computed using the remaining 226 (113 female and 133 male) identities. The ResNet50 was trained during 10 epochs and the final regression layer during 10 more epochs. Adam optimizer is used, with a learning rate of 0.01. The adopted loss function is Huber loss with $\delta = 1$.

Similarly to the setup presented in 3.3.2, we trained for HR estimation, a 3D-CNN for 10 epochs, on 96 videos from 24 subjects of the COHFACE dataset. Adam optimizer is selected with a learning rate set to 0.001 and a categorical cross-entropy loss function is used. The experiments are carried

[1]https://github.com/serengil/deepface
[2]https://github.com/arunponnusamy/cvlib
[3]https://github.com/siriusdemon/pytorch-DEX

out in our customized version of the testing set of COHFACE, i.e.: FFMF-COHFACE composed of 260 videos from 12 subjects not present at the training.

7.2.3 Results

In this Section, the performances are presented in tables where the highest values are indicated in **bold** text and the lowest values are underlined.

Gender Classification

In Table 7.1, the assessment of the two selected open-source gender classifiers is presented. Initially, when "original" images are passed to both estimators, we remark on some disparities in their performance for male and female subjects. *DeepFace*, although more reliable when tested on different datasets, performs to some degree better with the FFMF-CALFW images. Nevertheless, we can observe a significant disadvantage for the female subjects, indicating a potential bias towards this group. At first, *cvlib* does not suffer from this inequality but its performance is remarkably lower with the FFMF-CALFW faces in comparison with the subjects of the FFMF-VIP.

Regarding the "uploaded" images of FFMF-CALFW, we can notice some stability when compression is applied for both estimators while on the FFMF-VIP more variability can be observed. From the tables, we can see that *cvlib* is globally more robust than *DeepFace*, with less fluctuation in the results. However, when the face images are compressed we can observe a non-deterministic behavior of the gender networks. On the one hand, their performance can be slightly better as is the case for *cvlib* in the FFMF-CALFW and *DeepFace* in the FFMF-VIP for the male and female subjects, respectively, when the images are compressed through Snapchat and TikTok. On the other, *DeepFace*'s performance drops significantly for the female subjects when Instagram compression is applied on the FFMF-VIP faces. In further comparisons, we use the compressed images as our baseline since we aim to evaluate the filter impact alone, leaving aside other network disturbances such as compression.

When the filters are applied, the use of digital beautification enhances the ability of gender estimators to correctly classify female subjects. This is more explicit for *DeepFace*. We can see that for both datasets the male accuracy remains stable when filters are used. Exceptionally, for Relax! You Pretty! the accuracy for the male class drops. However, when filters are applied, the accuracy of the female group increases for almost all filters therefore the global accuracy augments accordingly. This trend is more evident for the

filters with a lower SSIM as shown in Table 6.2, such as the Snapchat filters and the Instagram filters Relax! You Pretty and Glam grain. Concerning *cvlib*, we remark fewer gender classification differences than *DeepFace* when non-filtered images were provided, nevertheless the observed impact of filters on this estimator is greater. Although the impact is minor for most of the filters, when Glam grain is applied to FFMF-CALFW and FFMF-VIP or when Mellow Glow or Belle are applied to FFMF-CALFW, not only the accuracy of the female subjects increases significantly but also the male class suffers a notable drop. When Spring glow is used, both classes suffer from a decrease in performance. In general, we observe how the most popular open-source gender estimators are not balanced by the classes they aim to predict and, similarly, the use of digital beautification via SN affects the female and male classes in different ways. While males are more difficult to recognize when beautification is applied, women benefit from this pre-processing.

By observing Table 6.1, we can see that filters enlarge the lips or shrink the nose size, modify the skin color and/or add some virtual makeup by including ARF, making their effect similar to the use of makeup, acting as foundation or lipstick or imitating some makeup techniques such as face contouring. Some works have associated the use of makeup with an increase of apparent femininity [161], by studying the impact of facial cosmetics on automated gender estimation, proving that it can negatively impact the employed neural networks, which then fail to classify men and women [82]. An analogous behavior is observed in our experiments where the use of filters increases the image femininity leading to an increase in female accuracy and a decrease in male recognition.

Age Estimation

Table 7.2 presents the results of the apparent age estimation experiments in the FFMF-VIP and FFMF-CALFW. When the images are "uploaded" to SNs, we observe small changes in the estimated age reported. When beautification filters are applied the correlation coefficient remains high, never going below 0.91. Even so, diverse effects can be observed. The study on the FFMF-VIP "beautified" dataset reveals that the application of almost all the filters modifies the apparent age estimation with similar strength. The computed MAE across all groups of images in the FFMF-VIP group ranges between 4.0 and 3.1 years for the female subjects and between 5.6 and 3.6 years for the male subjects. Nevertheless, the Instagram filter Thinner_face does not fall in this interval having close to no impact on the predicted age with a MAE of 0.8 and 1.1 years for the female and male classes, respectively. The computed MAE intervals are both wider and higher for the male subjects

Table 7.1: Assessment of the impact of beautification filters on *gender classification* on the FFMF dataset. Female (F), Male (M). The *higher* the value, the better.

		FFMF-CALFW						FFMF-VIP					
		DeepFace			cvlib			DeepFace			cvlib		
Experiment		**All**	**F**	**M**	**All**	**F**	**M**	**All**	**F**	**M**	**All**	**F**	**M**
Original		0.93	0.87	**0.99**	**0.81**	0.81	0.81	0.92	0.85	0.99	**0.96**	**0.98**	0.95
Ig	Uploaded	0.93	0.88	**0.99**	**0.81**	0.81	0.81	0.86	0.75	0.98	**0.96**	0.97	0.96
	Thinner_Face	0.92	0.85	**0.99**	**0.81**	0.80	0.81	0.87	0.76	0.98	**0.96**	0.97	0.95
	Relax! You Pretty!	0.94	0.90	**0.99**	0.80	0.82	0.78	0.87	0.79	0.96	**0.96**	0.97	0.94
	Glam Grain	0.95	0.91	**0.99**	0.80	**0.83**	0.77	0.94	0.88	0.99	**0.96**	**0.98**	0.94
Sc	Uploaded	0.93	0.88	**0.99**	0.80	0.79	0.82	0.94	0.89	0.99	**0.96**	0.97	0.95
	Fresh light	**0.96**	**0.93**	**0.99**	0.79	0.81	0.78	0.96	0.92	**1**	0.95	**0.98**	0.93
	Fresh vibes	0.95	0.91	**0.99**	0.79	0.81	0.78	0.96	0.94	0.99	0.94	0.95	0.94
	Mellow glow	0.95	0.92	**0.99**	0.79	**0.83**	0.76	**0.97**	**0.95**	0.99	0.94	0.99	0.90
Tk	Uploaded	0.93	0.88	**0.99**	**0.81**	0.80	0.83	0.95	0.90	0.99	**0.96**	0.96	0.95
	Belle	0.95	0.91	0.98	0.80	0.75	**0.85**	0.93	0.89	0.98	**0.96**	0.96	0.95
	Spring glow	**0.96**	**0.93**	**0.99**	0.74	0.74	0.74	0.95	0.90	0.99	0.94	0.97	0.92

indicating that the use of filters implies a higher disturbance when predicting the age of males. Moreover, we can observe an increase in the MAE and Mean for the male subjects with respect to the MAE and Mean of the female subjects when Spring glow, Relax! You pretty!, and Glam grain are applied.

On the FFMF-CALFW set, there are no big discrepancies between the estimated age from the "uploaded" and "beautified" images for the female and male subjects. The error metrics range between 3.1 and 0.8 years for the female subjects and between 3.3 and 0.8 years for the male subjects. A special case stands for the filter Spring glow delivering a MAE of 5.7 and 4.6 years for females and males respectively.

The presented discrepancies in terms of gender can be explained by some unwilling network bias induced at the training stage. Similar to the case of gender classification, the use of filters can be compared to the digital addition of makeup. In our experiments, we observed that the DEX estimator is more robust to disturbances induced by FFMF and ARF for female subjects than for the male category. This makeup variability is likely not present in the training set for male subjects, making this class more sensitive to filter disturbances.

Weight Estimation

In Table 7.3 the results of the weight network are presented. When predicting the weight from the "uploaded" images we observe that TikTok has the strongest impact on the weight model, increasing the network's MAE by

Table 7.2: Assessment of the impact of beautification filters on *apparent age estimation* on the FFMF dataset. The *lower* the Mean, StD and MAE, the better. The *higher* the ρ, the better.

CALFW	**Experiment**	**Female**					**Male**			
		Mean	StD	MAE	ρ		Me	StD	MAE	ρ
Ig	Uploaded	0.206	0.784	0.571	**0.997**		0.376	0.708	0.583	**0.998**
	Thinner_Face	0.461	1.149	0.879	0.994		0.612	0.961	0.840	0.996
	Relax! You Pretty!	1.302	1.996	1.764	0.982		1.402	1.809	1.770	0.989
	Glam Grain	1.924	2.252	2.188	0.979		1.796	1.847	2.023	0.988
Sc	Uploaded	0.328	1.159	0.887	0.994		0.391	0.975	0.778	0.996
	Fresh light	1.979	2.114	2.252	0.981		2.377	2.047	2.538	0.894
	Fresh vibes	3.025	2.682	3.177	0.969		3.23	2.455	3.350	0.978
	Mellow glow	2.586	3.008	2.984	0.962		3.234	2.756	3.360	0.976
Tk	Uploaded	0.318	1.254	0.900	0.993		0.235	1.004	0.755	0.996
	Belle	2.630	2.357	2.781	0.978		2.729	2.050	2.796	0.984
	Spring glow	**5.700**	**4.642**	**5.769**	0.915		**4.572**	**2.823**	**4.601**	0.971
VIP	**Experiment**	Mean	StD	MAE	ρ		Me	StD	MAE	ρ
Ig	Uploaded	0.558	1.002	0.804	**0.992**		0.634	0.936	0.859	**0.995**
	Thinner_Face	0.497	1.152	0.875	0.990		0.869	1.234	1.118	0.991
	Relax! You Pretty!	3.415	3.601	3.516	0.906		4.829	**3.701**	4.893	0.922
	Glam Grain	3.020	3.579	3.195	0.918		4.522	3.411	4.554	0.932
Sc	Uploaded	1.794	1.73	1.882	0.979		1.461	1.507	1.609	0.987
	Fresh light	3.416	3.182	3.467	0.931		3.881	2.740	3.930	0.957
	Fresh vibes	**4.025**	**3.762**	**4.097**	0.904		4.479	3.150	4.518	0.943
	Mellow glow	2.941	3.645	3.179	0.909		3.560	2.984	3.663	0.949
Tk	Uploaded	1.511	1.750	1.693	0.979		1.225	1.520	1.481	0.987
	Belle	3.952	3.596	4.002	0.914		4.667	3.027	4.681	0.948
	Spring glow	3.632	3.679	3.755	0.923		**5.617**	3.449	**5.622**	0.932

almost 0.5 kg while decreasing the ρ coefficient and the PAP metric by 0.3 and 18 points, respectively. The effect of Snapchat and Instagram filters is similar, increasing slightly the StD and MAE while dropping the PAP around 15%. Comparing "uploaded" with "beautified" images, we can observe stability in terms of StD, an increase in MAE ranging from 0.4 kg to 1.3 kg, a decrease of ρ up to 0.5 points and a decrease of PAP of a maximum of 8%. Glam Grain, Fresh vibes, Fresh light, Mellow glow, and Spring glow are the filters with a bigger impact on the MAE, making it drop of more than $0,5$ kg with respect to their corresponding SN "uploaded" images. As shown in Table 6.1, those filters modify biometric features such as nose and face contour, which are important for predicting the weight from face images being facial contour the most crucial region for this task, as presented in Section 3.2.4, being this feature modified by Fresh vibes, Fresh light, Mellow glow, Spring glow, and Thinner_face. The Me of the filter Thinner_face, the lowest among all

Table 7.3: Assessment of the impact of beautification filters on *weight estimation* on the FFMF-VIP. The *lower* the Me, StD and MAE, the better. The *higher* the ρ and PAP, the better.

	Experiment	**Me**	**StD**	**MAE**	ρ	**PAP**
	Original	2,79	22,82	8,52	**0,68**	**68 %**
Ig	Uploaded	2,55	22,94	8,65	0,67	54%
	Thinner_Face	2,42	**23**	8,69	0,67	54%
	Relax! You Pretty!	3,92	22,78	8,89	0,65	53%
	Glam Grain	2,86	22,42	9,16	0,64	47%
Sc	Uploaded	2,62	22,86	8,82	0,66	52%
	Fresh light	5,50	22,18	9,76	0,63	49%
	Fresh vibes	4,33	22,14	**9,85**	0,61	49%
	Mellow glow	4,83	22,27	9,83	0,61	44%
Tk	Uploaded	2,37	22,57	8,98	0,65	50%
	Belle	3,56	22,39	9,21	0,63	48%
	Spring glow	**5,52**	22,03	9,81	0,63	48%

categories of images, suggests that the weight estimator assigns lower weights for this type of data, which is consistent with the fact that the main effect of this filter is to shrink the face. The biggest drop in PAP with respect to the "uploaded" images is caused by Glam Grain and Mellow glow, filters that, as reported in Table 6.1, apply the highest amount of modifications to face images and, consequently, present some of the lowest SSIM on the FFMF-VIP as shown in Table 6.2.

Heart Rate Estimation

In Table 7.4, the evaluation of the HR estimation network on the videos of the FFMF dataset is presented.

The results displayed in Table 7.4 confirm our hypothesis, the application of beautification filters severely penalizes the HR estimation from face videos with higher StD and MAE values and lower correlation coefficient ρ between the real and the predicted HR when a filter is applied. More specifically, for the filters that apply skin smoothing, as presented in Table 6.1 Relax! You Pretty! and Hawaii_grain, the high StD values indicate a low model precision. Parallel to this metric, a value of $\rho < 0.40$ and $\rho < 0.20$ indicate low correlation and uncorrelation for the filters Relax! You Pretty! and Hawaii_grain, respectively. Regarding the filter Thinner_face, we can see a smaller increase of error than for the other filters studied which can be explained by the fact that this filter does not apply any skin smoothening as

Table 7.4: Assessment of the impact of SN filters on *HR estimation* on FFMF-COHFACE. The *lower* the SD and MAE, the better. The *higher* the ρ, the better.

Experiment	**SD**	**MAE**	ρ
Original	7.23	5.5	**0.62**
Uploaded	8.06	6.73	0.53
Thinner_face	8.67	6.31	0.51
Relax! You Pretty!	**10.98**	**9.1**	0.18
Hawaii_grain	10.12	8.32	0.38

shown in Table 6.1. We can also observe how the compression applied by the SN Instagram causes already a drop in the HR estimator performance.

As presented in Chapter 2, HR is a type of microsignal, defined as lower magnitude cues in strength or scale [5]. Some microsignals such as a person's HR might describe physical and behavioral attributes playing important roles in media security and forensics. Moreover, in privacy-preserving scenarios an end-user might desire that health-related information such as their HR, remains undetected in e.g. face recognition scenarios. As portrayed in Table 7.4, beautification filters prevent the HR from being successfully detected on face videos revealing themselves as an inexpensive way to protect those attributes and, consequently, the users' privacy.

7.3 Summary

In this Chapter, we have extended our study on the impact that the ongoing trend of digital face beautification through social media filters poses on facial processing tasks. The impact of each selected filter in the different studied tasks is summarized in Table 7.5. Filters that reported a low SSIM value, such as Glam Grain and Relax! You Pretty!, have consistently demonstrated a significant drop in the performance of the models in all experiments. Belle, the filter that shows a higher SSIM value, has been proven to act as light makeup thus increasing the user's feminity leading to an improvement in gender classification for females. However, filters undoubtedly act as a disturbance factor for AI-face facial biometric models.

In Chapters 6 and 7, we have highlighted that even easy-to-use social media filters can increase the likelihood of a fake video being classified as real, significantly increase the error on soft biometric estimators and can erase some microsignals. We underline that facial processing is not just a matter of having accurate systems in normal conditions but also understanding how

Table 7.5: Impact of each filter on the analyzed facial processing tasks. The filters are ranked from lower to higher aggressivity according to SSIM values, and the impact is measured by taking the compressed images as a baseline.

A	**Filter name**	**Gender**	**Age**	**Weight**	**HR**
G	Belle	++	++	+	-
G	Spring glow	++	+++	++	-
R	Thinner_Face	+	+	+	+
E	Fresh Lights	++	++	++	-
S	Mellow glow	+	++	+++	-
S	Fresh vibes	+	++	+++	-
I	Relax! Your Pretty!	++	++	++	+++
V	Glam Grain*	+++	++	+++	-
E	Hawaii Grain*	-	-	-	+++

+: Low impact, ++: Medium impact, +++: High impact. *Similar aggressivity.

common alterations can impact these systems. Retraining deep learning-based models with beautified data might not guarantee a solution, as filters are being created daily, making generalization difficult. Given the substantial impact of beautification filters, the use of a dedicated filter detection method is strongly advisable.

Part III

Conclusion and Author's Publications

Chapter 8

Conclusion and Future Directions

8.1 Conclusions

Daily estimation of biometric traits has become increasingly common in recent years. Examples such as automatic border controls, targeted marketing, and smartwatches illustrate how biometrics impact our habits and can enhance the quality of our lives. This dissertation explored the estimation of facial biometrics beyond people recognition, considering visible and thermal spectra and provided a quantified study of how current social media practices challenge existing facial processing networks. Some of the contributions of this thesis include:

Estimating biometric parameters that are not traditionally computed from facial input visuals

Extensive research on the estimation of soft biometrics, such as gender and age, has been performed in the literature. Their estimation from facial visuals has been proven, and several open-source methods exist for the average user to utilize.

Unlike other soft biometric traits, weight reflects both physical appearance and health status. Contrary to gender and height, body weight fluctuates throughout a person's adult life and requires regular monitoring. Moreover, weight estimation via a person's image has predominantly focused on full-body images and videos. In this thesis, we have presented several methods and training strategies for remote weight estimation from a single facial image, improving the accuracy of such methods and closing the gap between physical devices and deep learning structures for this trait estimation.

Exploring innovative deep-learning architectures

In this dissertation, we have not just improved the performance of specific networks for weight and heart rate estimation from face visuals but we have done pioneering research on the use of Vision Transformers for facial recognition tasks.

We have introduced a novel architecture for face embedding extraction: BEFiT. BEFiT is a vision transformer that can extract a universal face embedding without any prior knowledge of the facial processing targeted task from which different soft biometric traits can be estimated. Our experimental results demonstrate that this embedding achieves competitive performance with the ones extracted from dedicated architectures for gender, age and weight estimation.

Collecting the largest annotated paired visible-thermal dataset

The advancements in thermal imagery and its successful use for FR have strongly motivated the research done in this thesis. This dissertation presented the first dataset to provide paired visible-thermal face images and recordings with accompanying gender, age, body temperature, SpO2, BP, HR, height, weight, BMI, and 11 additional health metrics: the LVT Face Dataset for face biometrics. The extensive number of parameters annotated by every subject helped unlock the potential of thermal data for the estimation of three different soft biometric traits presented in this dissertation: gender, age, and weight.

Investigating the estimation of new parameters (gender age and weight) from face images in thermal spectra

Extensive experiments were conducted to demonstrate the feasibility of estimating three biometric traits: gender, age, and weight from facial thermal data. By partitioning the test set into three subsets based on the three variabilities presented in the LVT dataset, the results highlight the advantages of thermal imagery, especially for age and weight estimation from faces. They demonstrate that thermal imaging is superior when no dedicated light sources are used, such as in ambient light conditions.

Building on these promising results, thermal imagery was explored not only as an alternative but also as a complement to visible data. Using our novel Transformer architecture, we trained two different versions of BEFiT, and we fused BEFiT-V and BEFiT-T at the decision level. Our fusion strategy successfully estimates the three traits considered, achieving state-of-the-art performance for gender and age on the LVT dataset.

Analyzing the current trend of digital face beautification via social media filters and its impact on biometrics

We address the new trend of digital face beautification achieved through the application of social media filters, specifically examining its impact on the information recovered from filtered faces through deep learning structures. Our results demonstrate that with the use of beautification filters, the accuracy of security face verification systems can be significantly degraded, and the likelihood of a deepfake video being wrongly classified as authentic increases, thus challenging the reliability of those systems.

When more aggressive filters are used, the estimation of other traits such as age, weight, and HR becomes difficult, especially in the case of HR where skin smoothening techniques can delete microsignals. Despite more aggressive filters penalizing accuracy more heavily, the femininity score of the images is instead increased, leading to higher performance for female subjects and lower accuracy for male subjects in gender classification. This also causes female images to be less degraded in age estimation since makeup variation is likely to be present in the training set.

8.2 Directions for Future Research

Directions for future research relate to both the extension of the presented work to other facial image processing tasks, as well as the exploration of counter proposals for filter effects. Long-term perspectives include the utilization of new types of data beyond RGB and thermal for facial processing. Further work includes:

- **Transformers for video-based facial processing** Unlike static images, videos depict how objects move and interact over time. This dynamic nature brings additional difficulty for representation learning. We believe that the ability of Transformers to infer relations between long dependencies can be a powerful tool for video-based facial processing, such as heart rate estimation.

- **Thermal imagery for various health traits estimation.** The experiments presented in this thesis, highlighted the advantages of thermal imagery, especially for age and weight estimation from faces, and demonstrated that thermal imaging is superior to RGB data in some cases. The estimation of other parameters, such as heart rate, blood pressure, and oxygen saturation from thermal depictions can be of interest to the health and biometric community.

- **Filter Effect Countermeasures.** Filter removal can be considered, although much information might be lost. This could be of higher impact with the use of ARF where the recovered information might be visually correct, but any microsignals inserted could be fabricated. We advise including filtered images in the training step of biometric systems such as FR or gender estimation, as digital beautification is now a common practice in the real world. Another solution could be the use of a dedicated filter detection method that could trigger a warning to existing biometric systems.

- **Microsignal preservation.** Regarding microsignals such as HR, which can be erased from filtered videos, we propose two approaches. First, the design of beautification filters does not degrade hidden biometrics, allowing the machine to detect them while beautifying the video for the human eye. A proposed alternative is restoring the concealed micro signal after filtering the video.

- **Event data for microsignals extraction.** Unlike standard cameras that capture images at regular intervals, event cameras are motion-driven and respond to changes in the scene, capturing data only when pixels' intensities change. By processing the sparse, high-resolution data from event cameras, we believe that the deep learning model will be capable of accurately recognizing a broad spectrum of microsignals. Event data might have significant implications for areas where traditional RGB and thermal cameras struggle, such as in low-light conditions or dynamic real-world environments.

Author's Publications

Conferences

[C1] **N. Mirabet-Herranz**, M. Winter, Y. Lu, *et al.*, "Xaiface: A framework and toolkit for explainable face recognition," in *Proceedings of the 21th International Conference on Content-based Multimedia Indexing.*, 2024.

[C2] **N. Mirabet-Herranz**, C. Galdi, and J.-L. Dugelay, "One embedding to predict them all: Visible and thermal universal face representations for soft biometric estimation via vision transformers," in *IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, 2024, pp. 1500–1509.

[C3] **N. Mirabet-Herranz**, K. Mallat, and J.-L. Dugelay, "New insights on weight estimation from face images," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, IEEE, 2023, pp. 1–6.

[C4] **N. Mirabet-Herranz** and J.-L. Dugelay, "Lvt face database: A benchmark database for visible and hidden face biometrics," in *2023 International Conference of the Biometrics Special Interest Group (BIOSIG)*, IEEE, 2023, pp. 1–6.

[C5] **N. Mirabet-Herranz**, K. Mallat, and J.-L. Dugelay, "Deep learning for remote heart rate estimation: A reproducible and optimal state-of-the-art framework," in *International Conference on Pattern Recognition*, Springer, 2022, pp. 558–573.

[C6] **N. Mirabet-Herranz**, C. Galdi, and J.-L. Dugelay, "Impact of digital face beautification in biometrics," in *2022 10th European Workshop on Visual Information Processing (EUVIP)*, IEEE, 2022, pp. 1–6.

[C7] A. Libourel, S. Husseini, **N. Mirabet-Herranz**, and J.-L. Dugelay, "A case study on how beautification filters can fool deepfake detectors," in *IWBF 2024, 12th IEEE International Workshop on Biometrics and Forensics*, IEEE, 2024.

Journals

[J1] **N. Mirabet-Herranz** and J.-L. Dugelay, "Beyond the visible: Thermal data for facial soft biometric estimation," in *EURASIP Journal on Image and Video Processing*, 2024.

[J2] **N. Mirabet-Herranz**, C. Galdi, and J.-L. Dugelay, "Facial biometrics in the social media era: An in-depth analysis of the challenge posed by beautification filters.," in *IEEE Transactions on Biometrics, Behavior, and Identity Science (TBiom)*, 2024.

Prices

[P1] **N. Mirabet-Herranz**, "Social media filters: Beautification for humans but a critical issue for ai," *Three Minutes Thesis (3MT) contest at 31st European Signal Processing Conference (EUSIPCO)*, 2023.

Video journal

[V1] **N. Mirabet-Herranz**, "Social media filters: Beautification for humans but a critical issue for ai," *Science Talks*, 2024.

Datasets collected

[D1] N. Mirabet-Herranz, *Prisoners dataset*, `https://prisoners.eurecom.fr/`, 2023.

[D2] N. Mirabet-Herranz, *Label-eurecom visible and thermal face dataset*, `https://lvt.eurecom.fr/`, 2023.

[D3] N. Mirabet-Herranz, *Facial feature modification filters dataset*, `https://ffmf.eurecom.fr/`, 2022.

[D4] N. Mirabet-Herranz, *Prisoners dataset*, `https://celebdfb.eurecom.fr/`, 2024.

References

[1] K. Mallat and J.-L. Dugelay, "A benchmark database of visible and thermal paired face images across multiple variations," in *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*, IEEE, 2018, pp. 1–5.

[2] A. Dantcheva, C. Velardo, A. D'angelo, and J.-L. Dugelay, "Bag of soft biometrics for person identification: New trends and challenges," *Multimedia Tools and Applications*, vol. 51, pp. 739–777, 2011.

[3] A. Ross, S. Banerjee, and A. Chowdhury, "Deducing health cues from biometric data," *Computer Vision and Image Understanding*, vol. 221, p. 103 438, 2022.

[4] A. Nait-Ali, "Hidden biometrics: Towards using biosignals and biomedical images for security applications," in *International Workshop on Systems, Signal Processing and their Applications, WOSSPA*, IEEE, 2011, pp. 352–356.

[5] M. Wu, "Exploiting micro-signals for physiological forensics," in *Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security*, 2020, pp. 1–1.

[6] D. Anghelone, C. Chen, A. Ross, and A. Dantcheva, "Beyond the visible: A survey on cross-spectral face recognition," *arXiv preprint arXiv:2201.04435*, 2022.

[7] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[8] L. Sirovich and M. Kirby, "Low-dimensional procedure for the characterization of human faces," *Josa a*, vol. 4, no. 3, pp. 519–524, 1987.

[9] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition*, IEEE Computer Society, 1991, pp. 586–587.

[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, *Generative adversarial networks*, 2014. arXiv: 1406.2661 [stat.ML].

[11] J. Ho, A. Jain, and P. Abbeel, *Denoising diffusion probabilistic models*, 2020. arXiv: 2006.11239 [cs.LG].

[12] D. A. Reid, S. Samangooei, C. Chen, M. S. Nixon, and A. Ross, "Soft biometrics for surveillance: An overview," *Handbook of statistics*, vol. 31, pp. 327–352, 2013.

[13] S. Samangooei, B. Guo, and M. S. Nixon, "The use of semantic human description as a soft biometric," in *2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems*, IEEE, 2008, pp. 1–7.

[14] B. Hassan, E. Izquierdo, and T. Piatrik, "Soft biometrics: A survey: Benchmark analysis, open challenges and recommendations," *Multimedia Tools and Applications*, pp. 1–44, 2021.

[15] G. Antipov, S.-A. Berrani, and J.-L. Dugelay, "Minimalistic cnn-based ensemble model for gender prediction from face images," *Pattern recognition letters*, vol. 70, pp. 59–65, 2016.

[16] J. Mansanet, A. Albiol, and R. Paredes, "Local deep neural networks for gender recognition," *Pattern Recognition Letters*, vol. 70, pp. 80–86, 2016.

[17] S. Jia, T. Lansdall-Welfare, and N. Cristianini, "Gender classification by deep learning on millions of weakly labelled images," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, IEEE, 2016, pp. 462–467.

[18] A. D'Amelio, V. Cuculo, and S. Bursic, "Gender recognition in the wild with small sample size-a dictionary learning approach," in *International Symposium on Formal Methods*, Springer, 2019, pp. 162–169.

[19] Y. Fu, G. Guo, and T. S. Huang, "Age synthesis and estimation via faces: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 11, pp. 1955–1976, 2010.

[20] X. Wang, R. Guo, and C. Kambhamettu, "Deeply-learned feature for age estimation," in *2015 IEEE Winter Conference on Applications of Computer Vision*, IEEE, 2015, pp. 534–541.

[21] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output cnn for age estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4920–4928.

[22] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-variance loss for deep age estimation from a face," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5285–5294.

[23] R. Rothe, R. Timofte, and L. Van Gool, "Dex: Deep expectation of apparent age from a single image," in *Proceedings of the IEEE international conference on computer vision workshops*, 2015, pp. 10–15.

[24] K. Zhang, N. Liu, X. Yuan, *et al.*, "Fine-grained age estimation in the wild with attention lstm networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 3140–3152, 2019.

[25] X. Wang, R. Li, Y. Zhou, and C. Kambhamettu, "A study of convolutional sparse feature learning for human age estimate," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, IEEE, 2017, pp. 566–572.

[26] S. K. Gupta and N. Nain, "Single attribute and multi attribute facial gender and age estimation," *Multimedia Tools and Applications*, vol. 82, no. 1, pp. 1289–1311, 2023.

[27] C. Chen and A. Ross, "Evaluation of gender classification methods on thermal and near-infrared face images," in *2011 International Joint Conference on Biometrics (IJCB)*, IEEE, 2011, pp. 1–8.

[28] M. Abouelenien, V. Pérez-Rosas, R. Mihalcea, and M. Burzo, "Multimodal gender detection," in *Proceedings of the 19th ACM international conference on multimodal interaction*, 2017, pp. 302–311.

[29] N. Narang and T. Bourlai, "Gender and ethnicity classification using deep learning in heterogeneous face recognition," in *2016 International Conference on biometrics (ICB)*, IEEE, 2016, pp. 1–8.

[30] M. A. Farooq, H. Javidnia, and P. Corcoran, "Performance estimation of the state-of-the-art convolution neural networks for thermal images-based gender classification system," *Journal of Electronic Imaging*, vol. 29, no. 6, pp. 063 004–063 004, 2020.

[31] M. Abdrakhmanova, A. Kuzdeuov, S. Jarju, Y. Khassanov, M. Lewis, and H. A. Varol, "Speakingfaces: A large-scale multimodal dataset of voice commands with visual and thermal video streams," *Sensors*, vol. 21, no. 10, p. 3465, 2021.

[32] D. Han, J. Zhang, and S. Shan, "Leveraging auxiliary tasks for height and weight estimation by multi task learning," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, IEEE, 2020, pp. 1–7.

[33] M. L. Barr, G. Guo, S. E. Colby, and M. D. Olfert, "Detecting body mass index from a facial photograph in lifestyle intervention," *Technologies*, vol. 6, no. 3, p. 83, 2018.

[34] M. Jiang, Y. Shang, and G. Guo, "On visual bmi analysis from facial images," *Image and Vision Computing*, vol. 89, pp. 183–196, 2019.

[35] H. Siddiqui, A. Rattani, D. R. Kisku, and T. Dean, "Al-based bmi inference from facial images: An application to weight monitoring," in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2020, pp. 1101–1105.

[36] C. Velardo and J.-L. Dugelay, "Weight estimation from visual body appearance," in *2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, IEEE, 2010, pp. 1–6.

[37] D. Cao, C. Chen, D. Adjeroh, and A. Ross, "Predicting gender and weight from human metrology using a copula model," in *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, IEEE, 2012, pp. 162–169.

[38] R. D. Labati, A. Genovese, V. Piuri, and F. Scotti, "Weight estimation from frame sequences using computational intelligence techniques," in *2012 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSA) Proceedings*, IEEE, 2012, pp. 29–34.

[39] A. Dantcheva, F. Bremond, and P. Bilinski, "Show me your face and i will tell you your height, weight and body mass index," in *2018 24th International Conference on Pattern Recognition (ICPR)*, IEEE, 2018, pp. 3555–3560.

[40] M. Jain, S. Deb, and A. V. Subramanyam, "Face video based touchless blood pressure and heart rate estimation," in *2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2016, pp. 1–5.

[41] D. Yadav, N. Kohli, P. Pandey, R. Singh, M. Vatsa, and A. Noore, "Effect of illicit drug abuse on face recognition," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2016, pp. 1–7.

[42] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological measurement*, vol. 28, no. 3, 2007.

[43] S. Happy, A. Dantcheva, A. Das, F. Bremond, R. Zeghari, and P. Robert, "Apathy classification by exploiting task relatedness," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, IEEE, 2020, pp. 489–494.

[44] G. Giannakakis, M. R. Koujan, A. Roussos, and K. Marias, "Automatic stress detection evaluating models of facial action units," in *2020 15th IEEE international conference on automatic face and gesture recognition (FG 2020)*, IEEE, 2020, pp. 728–733.

[45] Y. Gurovich, Y. Hanani, O. Bar, *et al.*, "Identifying facial phenotypes of genetic disorders using deep learning," *Nature medicine*, vol. 25, no. 1, pp. 60–64, 2019.

[46] L. Basel-Vanagaite, L. Wolf, M. Orin, *et al.*, "Recognition of the cornelia de lange syndrome phenotype with facial dysmorphology novel analysis," *Clinical genetics*, vol. 89, no. 5, pp. 557–563, 2016.

[47] M.-Z. Poh, D. J. McDuff, and R. W. Picard, "Non-contact, automated cardiac pulse measurements using video imaging and blind source separation.," *Optics express*, vol. 18, no. 10, pp. 10 762–10 774, 2010.

[48] W. Verkruysse, L. O. Svaasand, and J. S. Nelson, "Remote plethysmographic imaging using ambient light.," *Optics express*, vol. 16, no. 26, 2008.

[49] G. Heusch, A. Anjos, and S. Marcel, "A reproducible study on remote heart rate measurement," *arXiv preprint arXiv:1709.00962*, 2017.

[50] W. Chen and D. McDuff, "Deepphys: Video-based physiological measurement using convolutional attention networks," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 349–365.

[51] X. Niu, H. Han, S. Shan, and X. Chen, "Synrhythm: Learning a deep heart rate estimator from general to specific," in *2018 24th international conference on pattern recognition (ICPR)*, IEEE, 2018, pp. 3580–3585.

[52] O. Perepelkina, M. Artemyev, M. Churikova, and M. Grinenko, "Heartrack: Convolutional neural network for remote video-based heart rate monitoring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 288–289.

[53] R. Song, S. Zhang, C. Li, Y. Zhang, J. Cheng, and X. Chen, "Heart rate estimation from facial videos using a spatiotemporal representation with convolutional neural networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, 2020.

[54] R. Špetlík, V. Franc, and J. Matas, "Visual heart rate estimation with convolutional neural network," in *Proceedings of the british machine vision conference, Newcastle, UK*, 2018, pp. 3–6.

[55] Z. Yu, X. Li, and G. Zhao, "Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks," in *30th British Machine Visison Conference: BMVC 2019. 9th-12th September 2019, Cardiff, UK*, The British Machine Vision Conference (BMVC), 2019.

[56] Z. Yu, W. Peng, X. Li, X. Hong, and G. Zhao, "Remote heart rate measurement from highly compressed facial videos: An end-to-end deep learning solution with video enhancement," in *Proceedings of the Int. Conference on Computer Vision*, 2019.

[57] L. Birla, S. Shukla, A. K. Gupta, and P. Gupta, "Alpine: Improving remote heart rate estimation using contrastive learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5029–5038.

[58] W. Sun, Q. Sun, H.-M. Sun, Q. Sun, and R.-S. Jia, "Vit-rppg: A vision transformer-based network for remote heart rate estimation," *Journal of Electronic Imaging*, vol. 32, no. 2, pp. 023 024–023 024, 2023.

[59] Y. Chen, J. Zhuang, B. Li, Y. Zhang, and X. Zheng, "Remote blood pressure estimation via the spatiotemporal mapping of facial videos," *Sensors*, vol. 23, no. 6, p. 2963, 2023.

[60] R. Takahashi, K. Ogawa-Ochiai, and N. Tsumura, "Non-contact method of blood pressure estimation using only facial video," *Artificial Life and Robotics*, vol. 25, pp. 343–350, 2020.

[61] A. Al-Naji, M. F. Mahmood, A. B. Fakhri, and J. Chahl, "Computer vision for non-contact blood pressure (bp): Preliminary results," in *AIP Conference Proceedings*, AIP Publishing, vol. 2804, 2023.

[62] W. Wang, Z. Wei, J. Yuan, Y. Fang, and Y. Zheng, "Non-contact heart rate estimation based on singular spectrum component reconstruction using low-rank matrix and autocorrelation," *Plos one*, vol. 17, no. 12, e0275544, 2022.

[63] K. S. Nair and S. Sarath, "Illumination invariant non-invasive heart rate and blood pressure estimation from facial thermal images using deep learning," in *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2021, pp. 1–7.

[64] Z.-K. Wang, Y. Kao, and C.-T. Hsu, "Vision-based heart rate estimation via a two-stream cnn," in *2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2019, pp. 3327–3331.

[65] X. Niu, S. Shan, H. Han, and X. Chen, "Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation," *IEEE Transactions on Image Processing*, vol. 29, 2019.

[66] F. Bousefsaf, A. Pruski, and C. Maaoui, "3d convolutional neural networks for remote pulse rate measurement and mapping from facial video," *Applied Sciences*, vol. 9, no. 20, 2019.

[67] B. Huang, C.-M. Chang, C.-L. Lin, W. Chen, C.-F. Juang, and X. Wu, "Visual heart rate estimation from facial video based on cnn," in *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, IEEE, 2020, pp. 1658–1662.

[68] X. Niu, X. Zhao, H. Han, *et al.*, "Robust remote heart rate estimation from face utilizing spatial-temporal attention," in *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, IEEE, 2019, pp. 1–8.

[69] B. Lokendra and G. Puneet, "And-rppg: A novel denoising-rppg network for improving remote heart rate estimation," *Computers in biology and medicine*, p. 105 146, 2021.

[70] M. Hu, F. Qian, X. Wang, L. He, D. Guo, and F. Ren, "Robust heart rate estimation with spatial–temporal attention network from facial videos," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 2, pp. 639–647, 2021.

[71] B. Li, P. Zhang, J. Peng, and H. Fu, "Non-contact ppg signal and heart rate estimation with multi-hierarchical convolutional network," *Pattern Recognition*, vol. 139, p. 109 421, 2023.

[72] T. Balaji, C. S. R. Annavarapu, and A. Bablani, "Machine learning algorithms for social media analysis: A survey," *Computer Science Review*, vol. 40, p. 100 395, 2021.

[73] A. Javornik, B. Marder, J. B. Barhorst, *et al.*, "'what lies behind the filter?'uncovering the motivations for using augmented reality (ar) face filters on social media and their effect on well-being," *Computers in Human Behavior*, vol. 128, p. 107 126, 2022.

[74] J. Yang, J. Fardouly, Y. Wang, and W. Shi, "Selfie-viewing and facial dissatisfaction among emerging adults: A moderated mediation model of appearance comparisons and self-objectification," *International Journal of Environmental Research and Public Health*, vol. 17, no. 2, 2020.

[75] C. Hogan, "Social media," *Friend Or Foe: Tackling the Issue of Social Media in Schools*, p. 1, 2022.

[76] S. Bhatt, "The big picture in the entire ar-filter craze," *The Economic Times*, 2020.

[77] C. Lavrence and C. Cambre, ""do i look like my selfie?": Filters and the digital-forensic gaze," *Social Media+ Society*, vol. 6, no. 4, p. 2 056 305 120 955 182, 2020.

[78] L. Ulrich, J.-L. Dugelay, E. Vezzetti, S. Moos, and F. Marcolin, "Perspective morphometric criteria for facial beauty and proportion assessment," *Applied Sciences*, vol. 10, no. 1, p. 8, 2019.

[79] A. Dantcheva and J.-L. Dugelay, "Assessment of female facial beauty based on anthropometric, non-permanent and acquisition characteristics," *Multimedia Tools and Applications*, vol. 74, pp. 11 331–11 355, 2015.

[80] S. Ueda and T. Koyama, "Influence of make-up on facial recognition," *Perception*, vol. 39, no. 2, pp. 260–264, 2010.

[81] C. Rathgeb, P. Drozdowski, D. Fischer, and C. Busch, "Vulnerability assessment and detection of makeup presentation attacks," in *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, IEEE, 2020, pp. 1–6.

[82] C. Chen, A. Dantcheva, and A. Ross, "Impact of facial cosmetics on automatic gender and age estimation algorithms," in *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, IEEE, vol. 2, 2014, pp. 182–190.

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

[84] M.-L. Eckert, N. Kose, and J.-L. Dugelay, "Facial cosmetics database and impact analysis on automatic face recognition," in *2013 IEEE 15th international workshop on multimedia signal processing (MMSP)*, IEEE, 2013, pp. 434–439.

[85] E. Derman, C. Galdi, and J.-L. Dugelay, "Integrating facial makeup detection into multimodal biometric user verification system," in *2017 5th International Workshop on Biometrics and Forensics (IWBF)*, IEEE, 2017, pp. 1–6.

[86] C. Rathgeb, A. Dantcheva, and C. Busch, "Impact and detection of facial beautification in face recognition: An overview," *IEEE Access*, vol. 7, pp. 152 667–152 678, 2019.

[87] R. Singh, M. Vatsa, H. S. Bhatt, S. Bharadwaj, A. Noore, and S. S. Nooreyezdan, "Plastic surgery: A new dimension to face recognition," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 3, pp. 441–448, 2010.

[88] C. Rathgeb, D. Dogan, F. Stockhardt, M. De Marsico, and C. Busch, "Plastic surgery: An obstacle for deep face recognition?" In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 806–807.

[89] S. Suri, A. Sankaran, M. Vatsa, and R. Singh, "On matching faces with alterations due to plastic surgery and disguise," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, IEEE, 2018, pp. 1–7.

[90] A. Bharati, M. Vatsa, R. Singh, K. W. Bowyer, and X. Tong, "Demography-based facial retouching detection using subclass supervised sparse autoencoder," in *2017 IEEE international joint conference on biometrics (IJCB)*, IEEE, 2017, pp. 474–482.

[91] N. Kose, N. Erdogmus, and J.-L. Dugelay, "Block based face recognition approach robust to nose alterations," in *2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, IEEE, 2012, pp. 121–126.

[92] T. Leyvand, D. Cohen-Or, G. Dror, and D. Lischinski, "Digital face beautification," in *ACM Siggraph 2006 Sketches*, 2006, 169–es.

[93] H. Chen, W. Li, X. Gao, and B. Xiao, "Aep-gan: Aesthetic enhanced perception generative adversarial network for asian facial beauty synthesis," *Applied Intelligence*, pp. 1–28, 2023.

[94] N. Diamant, D. Zadok, C. Baskin, E. Schwartz, and A. M. Bronstein, "Beholder-gan: Generation and beautification of facial images with conditioning on their beauty level," in *2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2019, pp. 739–743.

[95] Y. Zhou and Q. Xiao, "Gan-based facial attractiveness enhancement," *arXiv preprint arXiv:2006.02766*, 2020.

[96] S. Velusamy, R. Parihar, R. Kini, and A. Rege, "Fabsoften: Face beautification via dynamic skin smoothing, guided feathering, and texture restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 530–531.

[97] P. Hedman, V. Skepetzis, K. Hernandez-Diaz, J. Bigun, and F. Alonso-Fernandez, "On the effect of selfie beautification filters on face detection and recognition," *Pattern Recognition Letters*, vol. 163, pp. 104–111, 2022.

[98] C. Botezatu, M. Ibsen, C. Rathgeb, and C. Busch, "Fun selfie filters in face recognition: Impact assessment and removal," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 5, no. 1, pp. 91–104, 2022.

[99] P. Riccio, B. Psomas, F. Galati, F. Escolano, T. Hofmann, and N. Oliver, "Openfilter: A framework to democratize research access to social media ar filters," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12 491–12 503, 2022.

[100] S. Menon and A.-M. Kelly, "How accurate is weight estimation in the emergency department?" *Emergency Medicine Australasia*, vol. 17, no. 2, pp. 113–116, 2005.

[101] M. D. Bloice, P. M. Roth, and A. Holzinger, "Biomedical image augmentation using augmentor," *Bioinformatics*, vol. 35, no. 21, pp. 4522–4524, 2019.

[102] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[103] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[104] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5810–5818.

[105] J. Wang, L. Perez, *et al.*, "The effectiveness of data augmentation in image classification using deep learning," *Convolutional Neural Networks Vis. Recognit*, vol. 11, no. 2017, pp. 1–8, 2017.

[106] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, Springer, 2014, pp. 818–833.

[107] Q. Zhang, Y. Yang, H. Ma, and Y. N. Wu, "Interpreting cnns via decision trees," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6261–6270.

[108] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *Advances in neural information processing systems*, vol. 29, 2016.

[109] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[110] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[111] E. Hjelmas and J. Wroldsen, "Recognizing faces from the eyes only," in *Proceedings of the Scandinavian Conference on Image Analysis*, vol. 2, 1999, pp. 659–664.

[112] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[113] H. Proenca and J. C. Neves, "Soft biometrics: Globally coherent solutions for hair segmentation and style recognition based on hierarchical mrfs," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1637–1645, 2017.

[114] S. Kwon, J. Kim, D. Lee, and K. Park, "Roi analysis for remote photoplethysmography on facial video," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2015, pp. 4938–4941.

[115] E. Van Kampen and W. G. Zijlstra, "Determination of hemoglobin and its derivatives," *Advances in clinical chemistry*, vol. 8, 1966.

[116] X. Wang, W. Xie, and J. Song, "Learning spatiotemporal features with 3dcnn and convgru for video anomaly detection," in *2018 14th IEEE International Conference on Signal Processing (ICSP)*, IEEE, 2018, pp. 474–479.

[117] Y. Wang and A. Dantcheva, "A video is worth more than 1000 lies. comparing 3dcnn approaches for detecting deepfakes," in *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, IEEE, 2020, pp. 515–519.

[118] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 42–55, 2011.

[119] J. Hernandez-Ortega, J. Fierrez, A. Morales, and D. Diaz, "A comparative evaluation of heart rate estimation methods using face videos," in *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, IEEE, 2020, pp. 1438–1443.

[120] D. Anghelone, C. Chen, P. Faure, A. Ross, and A. 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)*, IEEE, 2021, pp. 1–8.

[121] M. J. Eddine and J.-L. Dugelay, "Gait3: An event-based, visible and thermal database for gait recognition," in *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*, IEEE, 2022, pp. 1–5.

[122] M. Rai, T. Maity, and R. Yadav, "Thermal imaging system and its real time applications: A survey," *Journal of Engineering Technology*, vol. 6, no. 2, pp. 290–303, 2017.

[123] A. Kuzdeuov, D. Koishigarina, D. Aubakirova, S. Abushakimova, and H. A. Varol, "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.

[124] X. Kevin and W. Bowyer, "Visible-light and infrared face recognition," in *Workshop on Multimodal User Authentication*, Citeseer, 2003, p. 48.

[125] S. Wang, Z. Liu, S. Lv, *et al.*, "A natural visible and infrared facial expression database for expression recognition and emotion inference," *IEEE Transactions on Multimedia*, vol. 12, no. 7, pp. 682–691, 2010.

[126] T. Gault and A. Farag, "A fully automatic method to extract the heart rate from thermal video," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2013, pp. 336–341.

[127] K. Panetta, Q. Wan, S. Agaian, *et al.*, "A comprehensive database for benchmarking imaging systems," *IEEE transactions on pattern analysis and machine intelligence*, vol. 42, no. 3, pp. 509–520, 2018.

[128] C. Barbosa Pereira, M. Czaplik, V. Blazek, S. Leonhardt, and D. Teichmann, "Monitoring of cardiorespiratory signals using thermal imaging: A pilot study on healthy human subjects," *Sensors*, vol. 18, no. 5, p. 1541, 2018.

[129] D. Poster, M. Thielke, R. Nguyen, *et al.*, "A large-scale, time-synchronized visible and thermal face dataset," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1559–1568.

[130] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR 2015)*, Computational and Biological Learning Society, 2015.

[131] D. Gyawali, P. Pokharel, A. Chauhan, and S. C. Shakya, "Age range estimation using mtcnn and vgg-face model," in *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IEEE, 2020, pp. 1–6.

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

[133] M. Morley, "Thermal conductivities of muscles, fats and bones," *International Journal of Food Science & Technology*, vol. 1, no. 4, pp. 303–311, 1966.

[134] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[135] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 104–12 113.

[136] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[137] H. Bao, L. Dong, S. Piao, and F. Wei, "Beit: Bert pre-training of image transformers," in *International Conference on Learning Representations*, 2021.

[138] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*, Springer, 2020, pp. 213–229.

[139] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8739–8748.

[140] Y. Zhong and W. Deng, "Face transformer for recognition," *arXiv preprint arXiv:2103.14803*, 2021.

[141] W. Su, Y. Wang, K. Li, P. Gao, and Y. Qiao, "Hybrid token transformer for deep face recognition," *Pattern Recognition*, vol. 139, p. 109 443, 2023.

[142] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[143] T. Zheng, W. Deng, and J. Hu, "Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments," *arXiv preprint arXiv:1708.08197*, 2017.

[144] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "Agedb: The first manually collected, in-the-wild age database," in *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 51–59.

[145] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[146] N. Bousnina, J. Ascenso, P. L. Correia, and F. Pereira, "Impact of conventional and ai-based image coding on ai-based face recognition performance," in *2022 10th European Workshop on Visual Information Processing (EUVIP)*, IEEE, 2022, pp. 1–6.

[147] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3207–3216.

[148] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 225–14 234.

[149] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, and Z. Ge, "Implicit identity leakage: The stumbling block to improving deepfake detection generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3994–4004.

[150] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.

[151] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4113–4122.

[152] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[153] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 044–15 054.

[154] "Information technology — Biometric performance testing and reporting — Part 1: Principles and framework," International Organization for Standardization/International Electrotechnical Commission, Standard, May 2021.

[155] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "Qualitycrowd—a framework for crowd-based quality evaluation," in *2012 Picture coding symposium*, IEEE, 2012, pp. 245–248.

[156] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*, Springer, 2016, pp. 87–102.

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

[158] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.

[159] P. Korshunov and S. Marcel, "Subjective and objective evaluation of deepfake videos," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 2510–2514.

[160] V. Blazek, "Ambient and unobtrusive cardiorespiratory monitoring," in *2016 ELEKTRO*, IEEE, 2016, pp. 2–2.

[161] M. Carter, "Facials: The aesthetics of cosmetics and makeup," vol. 8, 1998.