On Designing Deep Learning-based Multispectral Imaging Systems: From Face Recognition to Firearm Detection

by

Jacob Rose

(Under the Direction of Thirimachos Bourlai)

Abstract

Deep learning-based systems have shown superior performance in many artificial intelligence tasks in the last decade. Imaging systems are particularly well suited to benefit from deep learning with the use of Convolutional Neural Networks (CNNs). CNNs and other deep learning methods are now the preferred approach for nearly all computer vision and pattern recognition challenges, including image classification, object detection, and biometric recognition. Not only are CNNs used with images captured in the visible spectrum wavelengths, but they have also been used with varying success on images captured in other wavelengths, most often in the infrared (IR) spectrum. The different regions of the IR band each have their own unique advantages, providing researchers with new ways to investigate computer vision problems in challenging scenarios that may be impossible to solve in the visible spectrum. Although deep learning-based imaging systems have been used successfully for a variety of tasks in both the visible and IR spectrums, many challenges remain, especially for biometric recognition applications. Studies show that a performance gap exists between the image-based systems that operate in different spectrums. These challenges can generally be attributed to the large appearance variations of images captured in the different spectrums and a limited availability of high quality data in the non-visible bands.

In this dissertation I address several of the gaps that exist in the open literature for imaging systems that use multiple spectrums, particularly as they relate to biometric systems and object detection. The dissertation includes an introduction and background information for each of the challenges covered in this work and a systematic review of the relevant literature. Methods are

proposed to address facial attribute analysis in the visible and middle-wave IR bands, facial landmark detection and recognition via image synthesis using the visible and passive (middle and long wave IR) bands, and detection of firearms from surveillance videos in the visible band. The proposed approaches and results from this dissertation provide practical solutions and analysis for a variety of imaging system challenges and provide helpful insights and directions for future research.

INDEX WORDS: Convolutional neural networks, deep learning, biometrics, face attributes, face landmark detection, generative adversarial networks, image synthesis, thermal imaging, face mask classification, multispectral imaging, face dataset, firearm detection, object detection

On Designing Deep Learning-based Multispectral Imaging Systems: From Face Recognition to Firearm Detection

by

Jacob Rose

B.S., West Virginia University, 2015

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

Doctor of Philosophy

Athens, Georgia

2023

On Designing Deep Learning-based Multispectral Imaging Systems: From Face Recognition to Firearm Detection

by

Jacob Rose

Major Professor:	Thirimachos Bourlai
Committee:	Jin Ye
	Kyle Johnsen
	Mable Fok

Electronic Version Approved:

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

Dedication

I dedicate this dissertation to my parents, who have always encouraged and supported me.

Acknowledgments

Thank you to my advisor and mentor, Dr. Thirimachos Bourlai, for his guidance and encouragement. I also wish to thank my family, friends, and teachers for their time, advice, and moral support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem Definition

In the last decade, deep learning has become the most popular machine learning method for solving complex pattern recognition and computer vision problems. One of the fields that has benefited greatly from deep learning is biometrics, a subfield of computer vision. Biometrics is the science of identification and verification of an individual based on the behavioral and physiological traits of the person. The traits are unique, permanent, and can separate one individual from another Dargan and Kumar, 2020. Biometric recognition can be considered as a verification or identification problem. In the context of face recognition, verification is a 1:1 matching problem where a presented face is confirmed or repudiated by comparing it with the claimed identity in the database. Identification is a 1:N matching problem where an unknown face is compared with each face in a database of already known identities and a decision is made with the comparisons Taskiran et al., 2020. Biometric systems are commonly used with sensors that capture images in the visible spectrum, but they can also be used with sensors in other spectra, including the infrared (IR) spectrum.

This dissertation presents work on several aspects of deep learning-based imaging systems in multiple spectrums. The main focus is on biometric systems. Most biometric systems are composed of three basic modules; (1) an acquisition and preprocessing module where data is collected and any required attributes in the data are detected and analysed. These attributes, often called soft biometrics Jain et al., 2004, are normally not unique enough to identify specific individuals but can be used as a compliment to the primary biometric modality; (2) a keypoint detection and normalization module where important landmarks are located, if necessary, and required alignment and image normalization processes are performed. Each of these processes is dependent upon the biometric

modality being used; (3) a feature extraction and matching module, where information is extracted from the image and similarity scores are obtained between biometric samples using an appropriate metric.

In this work I also focus on object detection, specifically firearm detection in surveillance images. The detection of objects in images and videos is one of the central and most challenging problems in computer vision. Given an image, the generic object detection task is to determine whether or not there are instances of objects from predefined categories and, if present, return the spacial location and extent of those instances L. Liu et al., 2020. Previously, object detection performance with handcrafted features reached a plateau after 2010 , but the use of CNNs like AlexNet Krizhevsky et al., 2012 in 2012 allowed new object detectors to learn robust and high-level feature representations of images Zou et al., 2023. Advances in object detection have also lead to researchers addressing more challenging objects to detect, such as firearms. The automatic detection of firearms is very difficult due to variations in size, shape, and appearance. Several algorithms have been proposed over the last few years to overcome these challenges Iqbal et al., 2021; Yadav et al., 2022, focusing mainly on accuracy and speed of detections. I contribute to this important research with an analysis of which data augmentations improve and hinder firearm detection on a novel dataset composed of only real-world surveillance images.

I present my contributions for each of the mentioned biometric modules and object detection in the next sections. The remainder of the paper is organized as follows: Chapter 2 reviews previous works related to multispectral imaging systems. Chapter 3 describes the methodology for each problem. The experimental results are discussed in Chapter 4 and conclusions are drawn in Chapter 5.

1.2 Acquisition and Preprocessing

1.2.1 Facial Attribute Analysis: Mugshot Data

In this section, I provide a brief introduction to forensic biometrics and introduce a forensic toolkit for analysis of a variety of facial attributes that affect face recognition. The contributions of this work are the following:

- Rapidly categorize face databases for factors that can degrade face recognition accuracy.

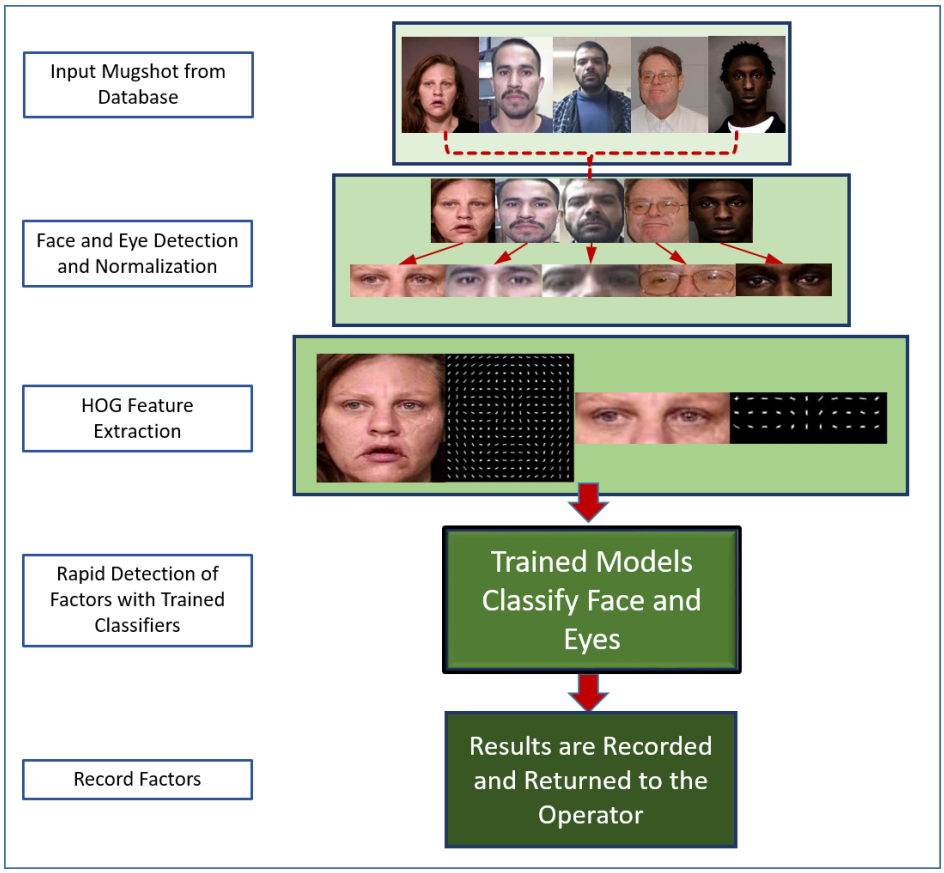


Figure 1.1: Overview of the toolkit work flow. Step 1: a mugshot image is imported into the interface. Step 2: the face and eye pair (if possible) are detected. Step 3: HOG features are extracted from the detected face and eyes. Step 4: the features are used for classification. Step 5: The classification results are recorded and returned to the examiner.

- Categorize the following factors (1) are the subject's eyes open or closed, (2) is the subject wearing glasses, and (3) is the facial pose of the subject frontal or non-frontal.
- Train classical and deep learning classification models that are robust to factors such as pose, illumination, expression, and resolution in high-quality face datasets collected under ideal conditions and low-quality mugshot face datasets collected under variable conditions.

The goal of both biometric recognition and forensic science is to link biological data to an individual Jain and Ross, 2015. However, the ability to use biometric systems successfully in forensic scenarios is quite challenging. The

challenges in this field, often referred to as forensic biometrics, as well as their similarities and differences are well documented in Champod and Tistarelli, 2017; Dessimoz and Champod, 2008; Jain et al., 2011, 2012; Jain and Ross, 2015; Lee et al., 2012; Meuwly and Veldhuis, 2012; Tistarelli et al., 2014. According to Meuwly and Veldhuis, 2012, biometric technology plays a role in several forensic applications: the identity management and the identity verification in the criminal justice chain, the identification of missing persons from a mass disaster, and the forensic investigation and intelligence as well as the forensic evaluation of biometric evidence in court, which together form the field of forensic biometrics. More specifically, and explained in Dessimoz and Champod, 2008, forensic biometric systems are used as sorting tools which do not make any final identification decisions. For forensic face recognition scenarios, an unknown probe face image is compared to every other face image in a gallery database. The FR system computes a similarity score for the probe with each sample in the gallery and the top-K matches are returned, often ordered from most to least similar. Then the forensic investigator performs a visual inspection of each candidate from the list to determine if any of the returned faces are a match to the unknown probe, meaning that the forensic biometric system is an external tool from the manual identification process.

Some of the major challenges in unconstrained face recognition are variations in pose, expression, occlusion, age, and image quality factors such as illumination, blurriness, and brightness. To improve face recognition performance it is important to identify which images in a database have these attributes so that they may be further analyzed or enhanced. The three factors I focus on are (1) whether a subject's eyes are open or closed, (2) whether the subject is wearing glasses or not, and (3) whether the facial pose of the subject is either frontal or non-frontal.

The goal of this work is to help the forensic operator by improving the process of returning an accurate rank list of potential suspects. I propose a toolkit that can rapidly categorize large databases for several factors that can degrade FR accuracy by detecting facial photos where the subject's eyes are closed, the subject is wearing glasses, or has a non-frontal face pose. The ability to identify these attributes from facial photos in a large database can benefit law enforcement and give operators the option to exclude, group, or enhance these images. An overview of the proposed system is presented in figure 1.1 where the system input is a mugshot or other face image from the database to be categorized. Then, face detection is performed as well as eye pair detection, if possible. Next, HOG features are extracted from the detected face and eyes

and each of the three factors are categorized by trained classifiers. The results are then recorded and available for analysis by an examiner.

1.2.2 Facial Attribute Analysis: Cellphone Data

The facial attribute analysis work from the previous section is expanded upon here. The contributions of this work are the following:

- Rapidly categorize face databases for factors that can degrade face recognition accuracy.
- Categorize the following factors (1) are the subject's eyes open or closed, (2) is the subject wearing glasses, and (3) is the facial pose of the subject frontal or non-frontal.
- Train classical and deep learning classification models on challenging face images collected using an iPhone 5S indoors and outdoors at distances of 1 and 5 meters.

Facial attributes, sometimes called soft biometrics, can be used to support human recognition systems in law enforcement scenarios where the focus is to reduce the search space and retrieve more relevant results to the query face images Martin et al., 2016. Soft biometrics can also be fused with hard biometrics to improve the efficiency of biometric systems Gonzalez-Sosa et al., 2018.

There are two main facial attribute analysis processes. The first one is Facial Attribute Manipulation (FAM), in which generative models are used to modify face images in order to alter or even remove a chosen attribute. I wil not cover the FAM based methods, but more information can be found in Shen and Liu, 2017; Xiao et al., 2018.

The second process, which I focus on, is Facial Attribute Estimation (FAE), in which specific classifiers are used to determine whether a particular facial attribute is present in a query face image Zheng et al., 2018. While there are several techniques proposed claiming to efficiently estimate different facial attributes, many challenges have yet to be solved related to the performance of face recognition systems. Thus, while common factors that affect face recognition systems include face pose, illumination, and occlusion, FAE challenges can be summarized by three different aspects: data, algorithms, and applications Zheng et al., 2018. Here, I am interested in using FAE processes to detect certain facial attributes so that higher recognition performance can be achieved. This is a difficult task, especially when working with a large amount of diverse, class imbalanced, and attribute specific datasets. For instance, when trying to classify

whether or not a face has eyeglasses present, most publicly available face datasets have far more face samples where the subjects are not wearing eyeglasses. This problem can lead to overfitting and poor performance on new data. To overcome this problem, data augmentation is often used to create more samples of the minority class in order to correct the imbalance and avoid bias. In addition to class imbalance, the source and distributions of data can affect a classifiers ability to generalize to data from sources not seen during training. Exploring the effects of the source and distributions of the data in object detection and recognition performance is very important.

I address the previously mentioned challenges using data captured from traditional cameras and mobile devices, when operating at multiple standoff distances, in indoor and outdoor conditions. I propose an approach that automatically and efficiently detects three specific facial attributes: (1) determining whether the eyes of a subject are open or closed, (2) determining whether a subject is wearing glasses or not, and (3) detecting whether a subject's facial pose is either frontal or non-frontal. To detect all of the aforementioned facial attributes, I trained and tested classical and deep learning based models. First, I assessed the previously trained classifiers from my work in 5.1.1 using traditional and deep learning methods on data collected with an iPhone 5S. Then, I retrained the classifiers on iPhone data to assess the performance changes. Finally, I retrained on a face database containing the previous data and mobile data combined. The proposed attribute-specific detection models are robust, yielding up to 100% accuracy (in terms of F1 score) depending on the attribute tested.

1.2.3 Facial Attribute Analysis: Masked Data

Facial attribute Analysis in the visible and mid-wave spectrums is investigated in this work during the COVID-19 pandemic, specifically focusing on deep learning methods for the classification of mask compliance. The main contributions are the following:

- The creation of a multispectral masked face (MMF-DB) database of 100 subjects with various levels of non-compliant and compliant mask wearing in the visible and middle wave infrared (MWIR) bands.
- The augmentation the MMF-DB with synthetically applied masks at two levels of non-compliance, masks placed below the nose and below the mouth.
- The assessment of the performance of nine well established CNN architectures on masked and unmasked face images.

- The development of an efficient deep learning based approach on solving the problem of classifying face images wearing masks as either compliant or non-compliant when operating in either the visible or thermal bands. Experimental results show that face mask compliance classification in both studied bands yield a classification accuracy that reaches 100% for most models studied, when experimenting on frontal face images captured at short distances with adequate illumination.

During the COVID-19 pandemic, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the World Health Organization (WHO) reported that wearing face masks helps prevent our respiratory droplets from reaching others. In addition, when wearing a face mask over the nose and mouth, the spray of droplets is reduced. In response to this recommendation, different governments around the world started a set of initiatives, including ones that aimed to utilize machine learning techniques to detect whether passengers are wearing face masks in metro stations. Many other members of the artificial intelligence community also started developing various automatic face mask detection models that can aid in the monitoring and screening of face mask usage. However, most reported models typically focused on detecting whether a mask is present in a face image or not. According to the mask guidelines provided by the Center for Disease Control and Prevention (CDC), face masks should (1) have two or more layers of washable, breathable fabric, (2) completely cover the nose and mouth, (3) fit snugly against the sides of the face with no gaps, and (4) have a nose wire to prevent air from leaking out the top of the mask. Therefore, the methods that only detect the presence of a mask will fail to identify subjects who are improperly wearing their mask and thus, not complying with CDC guidelines. It is important to note that in this work I am using 2D face images, and I focus on detecting whether the guideline (2) discussed above is being followed or not. More specifically, I want to determine if a detected face is "compliant", where a face mask is properly worn over the nose and mouth, or "non-compliant", where a detected face has either, (a) no face mask, (b) a mask worn below the nose, or (c) a mask worn below the chin. While I cannot account for all types of noncompliance, especially if the mask fits snugly against the sides of the face without gaps, scenarios (b) and (c) seem to be the most common cases of masked non-compliance observed during the pandemic. In this work, when I mention compliance with face mask guidelines, I will be specifically talking about the automation of monitoring and detecting compliance of wearing face masks that properly cover the nose and mouth areas.

Although there are several mask detection works proposed since the start of the pandemic, very few address different levels of face mask non-compliance,

including (a) no face mask present, (b) face mask is below the nose, and (c) face mask is below the mouth and nose. To my knowledge, only one large scale publicly available dataset exists with visible band face images annotated for masks that are present but not worn correctly Batagelj et al., 2021. I have found no work extending face mask compliance in the thermal band, specifically, MWIR. MWIR sensors operate on the passive IR band in the 3.0-5.0 μm spectral range. The benefits of operating in the MWIR band are numerous, with many applications including biometrics and biomedical Bourlai, 2016; Bourlai and Hornak, 2016; Bourlai, Pryor, et al., 2012; Bourlai, Ross, et al., 2012; Mokalla and Bourlai, 2019, 2020; Osia and Bourlai, 2012, 2017a. Passive IR sensors need no external light source, and instead detect IR radiation as heat emitted from objects. In addition to being tolerant of other commonly encountered environmental conditions such as fog, smoke, and dust, MWIR imaging sensors are ideal when operating under low light, night-time environments. Any operational scenario with less-than-ideal lighting conditions can greatly benefit from the use of MWIR imaging sensors.

In this work, I focus on classifying faces wearing masks as either compliant (mask properly covers the nose and mouth face areas) or non-compliant (not wearing a face mask or wearing one but it is placed below the nose or mouth). In addition to evaluating classification performance on visible band face images, I also investigate how different CNN architectures perform on face images where subjects are wearing face masks captured with an MWIR imaging sensor.

1.2.4 MultiSpectral Face Dataset Collection

In this section I introduce the data collection activities that I was a part of that used the latest MWIR imaging sensors, with and without telephoto capabilities, i.e., the A8581 (for close-range) and the RS8500 series (for long-range imaging). This work created the MILAB-VTF(B) dataset; a MWIR-Visible face dataset from which a curated version will become publicly available. Thus, assisting the research community by further closing the gap of MWIR-Visible datasets availability. The contributions of this work are the following:

- An unconstrained, multispectral (visible and MWIR), unsynchronized, paired face dataset with 400 identities.
- Challenging variations in terms of weather, pose, and distances.
- The largest multispectral face dataset to date by number of subjects, distances, and thermal sensor resolution.

The raw dataset was collected within the 1st quarter of 2021, and is the largest and most diverse of its kind to-date, collected in realistic operational conditions, from 1.5 - 400 meters (1312 ft). In the next chapters I will discuss more of the data collection activities as well as the demographics of the dataset. All work in this dissertation related to the dataset collection will soon be published in a journal that is accepted in IEEE Transactions on Biometrics, Behavior, and Identity Science.

1.3 Keypoint Detection

1.3.1 Facial Landmark Detection

Detecting landmarks in face images is an important preprocessing step for a variety of biometric and human physiology-related artificial intelligence applications, including head pose estimation, driver drowsiness detection, gaze tracking, emotion recognition, and face recognition. Face recognition in particular is one of the most studied problems in the biometric literature because of its many applications in security, surveillance, authentication, and identification tasks. In the visible light domain, the accurate detection of facial landmarks and alignment to canonical coordinates is a requirement for many face recognition algorithms. If face recognition in low-light or nighttime environments is the operation scenario of interest, thermal imaging cameras are often used due to their ability to passively capture body heat emissions without being dependent on the presence of or intensity level of ambient light conditions. However, facial landmark detection algorithms made for the visible domain are not designed to perform well when operating in the thermal domain. In order to accurately detect landmarks in thermal face images, spectral-dependent approaches must be developed. The contributions of this work are the following:

- Provide a comprehensive assessment of synthesis and transfer learning facial landmark detection approaches on the ARL-VTF and MILAB-VTF(B) datasets.
- Conduct a study where the focus is to train and evaluate HRNet facial landmark detectors on thermal and multispectral (thermal and visible) data. The purpose of such a study is to identify any performance gaps that exist between learning the invariant features between two domains and fine-tuning a model to the target domain with the chosen HRNet architecture.

- Propose a competitive facial landmark detection approach that yields state-of-the-art results on the ARL-VTF and the recently developed MILAB-VTF(B) datasets. Results are reported on face images captured at 1.5-, 100-, 200-, 300-, and 400-meter distances respectively.

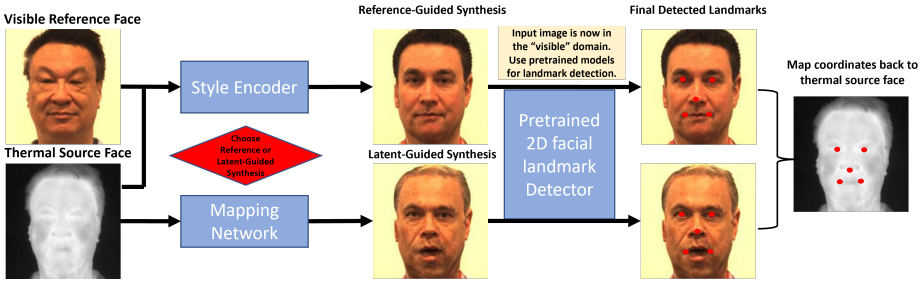


Figure 1.2: Overview of the reference-guided and latent-guided thermal-to-visible synthesis process using StarGAN v2. The thermal source face is the input image for both synthesis methods. StarGAN v2 uses the style encoder for reference-guided synthesis to output a visible image that looks similar to the reference image. The mapping network performs latent-guided synthesis to output a visible image using the latent codes learned from the dataset during training.

Conventional facial landmark detection methods can be grouped into three categories, Constrained Local Model (CLM) methods, holistic methods, and regression-based methods Y. Wu and Ji, 2019. Many of these methods work well on datasets collected under controlled conditions where there is little to no variation in head pose, illumination, expression, and occlusions. However, they often fail to accurately detect facial landmarks on "in-the-wild" scenarios that are composed of many, or all, of the aforementioned challenges. Recently developed algorithms based on Convolution Neural Networks (CNNs) and deep learning have proven much more suited to handling the problem of unconstrained in-the-wild facial landmark detection. These neural network algorithms can be grouped into two categories:

1. Coordinate regression methods where the model predicts (x, y) coordinates for every landmark.

2. Heatmap-based regression methods that use 2D heatmaps for every landmark Khabarlak and Koriashkina, 2021. Each heatmap is a 2D Gaussian kernel, often with the same standard deviation for all landmarks. In this approach, the values in the heatmap can be thought of as probabilities of the landmark location. Overall, the heatmap-based algorithms are generally more accurate on many of the common benchmark datasets.

The coordinate regression and heatmap-based regression methods can both be used in the thermal domain but require additional techniques to be implemented in order to achieve accurate detections as they are primarily developed using visible domain data. The two most commonly used techniques are transfer learning Zhuang et al., 2020 and image synthesis L. Wang et al., 2020. Transfer learning is widely used in many computer vision tasks. It fine-tunes previously trained networks to perform new tasks when a large scale labeled dataset is not available to train the network from scratch. The goal of transfer learning is to use weights learned from a model trained on a source domain and translate it to the desired target domain. The other method that has gained interest in the recent literature is the implementation of Generative Adversarial Networks (GANs) Goodfellow et al., 2020 to synthesize new data.

Generative models for image synthesis have become very popular for computer vision applications, and are one of many methods for face recognition in the thermal domain Bourlai, 2016; C. Chen and Ross, 2019; Di et al., 2021; Di et al., 2018; Fu et al., 2021; Osia and Bourlai, 2017a, 2017b; Osia et al., 2018; Peri et al., 2021a; Z. Wang et al., 2018; H. Zhang et al., 2019. Synthesis approaches use generative models to transform images from a source domain to a target domain. The source domain often lacks sufficient amounts of data for training a model using traditional methods. The target domain is usually well studied and has an adequate amount of labeled data. Several successful techniques for solving problems in this domain already exist. In this case, the source domain is face images in the thermal infrared band, and the target domain is face images in the visible band, see figure 1.2. The main challenge researchers face in this area, is that many accurate facial landmark detectors have been developed for the visible domain, largely due to the quality, size, and availability of visible face landmark detection datasets Burgos-Artizzu et al., 2013; Koestinger et al., 2011; Sagonas et al., 2013; W. Wu et al., 2018. In contrast, the thermal domain lacks face datasets with the diversity and size of the visible domain, making it a very challenging problem to solve. In this work, instead of creating new datasets with a sufficient number of samples for training new models, it is arguably easier to leverage already existing methods in the visible domain and apply them to the thermal domain using synthesis-based techniques.

There are several methods developed for facial landmark detection in the thermal domain Bourlai and Jafri, 2011; Chu and Liu, 2019; Kopaczka et al., 2018; D. Poster et al., 2019; Riggan et al., 2018, which can then be used to perform face recognition Anghelone et al., 2022. The methods that use transfer learning on thermal datasets often lack a large enough number of subjects and images to approach results reported in the visible domain. Additionally, thermal

images frequently have lower resolutions than those in the visible domain. This can largely be attributed to the higher cost of thermal imaging sensors compared to visible ones. Training landmark detection models solely on thermal data via transfer learning is most often used. However, thermal-to-visible synthesis using GANs has also proven to be a suitable approach. While both methods can produce satisfactory results, there is discussion in the literature that there is room for improvement. Synthesis methods are relatively new and are often difficult to train Gonog and Zhou, 2019; Gui et al., 2021; Roth et al., 2017. Also, many early synthesis models such as Osia and Bourlai, 2017a required paired data, but there are relatively few face datasets with paired images. Those that do have an inadequate number of images to train highly accurate models. With the introduction of CycleGan Zhu et al., 2017, synthesis techniques were no longer constrained to paired data. For this work, I choose methods based on the CycleGAN principles that can preserve the thermal domain source characteristics of a face, i.e., pose, location of the eyes, mouth, and so on, while generating faces that are similar to those in the target visible domain without the requirement of paired face image data.

Recent advances in synthesis methods, the improved image quality of new thermal imaging sensors, as well as new datasets such as the ARL-VTF D. Poster et al., 2021 and MILAB-VTF(B) Peri et al., 2021a are helping to advance the state-of-the-art in thermal facial landmark detection. However, a performance gap remains between results in the thermal and visible domains. In this work I implement recently developed methods for image synthesis and facial landmark detection. I use StarGAN v2 Choi et al., 2020 and CUT Park et al., 2020 for image synthesis and HRNet J. Wang et al., 2020 for facial landmark detection. I propose to find the strengths and limitations of these methods and determine whether current synthesis methods are able to outperform transfer learning techniques for thermal facial landmark detection.

1.4 Feature Extraction and Matching

1.4.1 Race and Gender Classification for Cross-spectral Face Recognition

Face recognition (FR) in the visible and MWIR spectrums using thermal-to-visible reference-guided synthesis is investigated in this work. The synthesis process is aided by classification of gender and race for face images in the MWIR spectrum in order to better preserve the discriminative features of the faces. The main contributions are the following:

- A multitask learning approach for predicting gender and race from MWIR face images based on the VGG Simonyan and Zisserman, 2014 architecture.
- An approach for cross-spectral face recognition that implements soft biometrics to improve thermal-to-visible synthesis of face images in order to reuse state-of-the-art face matchers.
- A comprehensive set of experiments investigating the effect of choosing the wrong gender or race when selecting a face image using thermal-to-visible reference-guided synthesis for cross-spectral face recognition.
- Superior face recognition performance using my proposed method for race and gender classification when compared to selecting a random gender or race.
- Baseline results for gender and race classification on the MILAB-VTF(B) dataset.

Face recognition is one of the most important challenges in computer vision. Face recognition is used in daily in many application and industries including law enforcement, surveillance and security systems, entertainment, shopping, and finance. Deep learning methods have been able to achieve human level recognition performance or better since 2014 Taigman et al., 2014. Recent face recognition algorithms perform very well on visible spectrum images collected under controlled conditions. However, face recognition in low light and night-time environments is far from being a solved problem. Methods developed for face recognition in day time conditions do not work well when illumination conditions are not ideal. Therefore, new methods for these unconstrained environments must be developed.

In order to use face recognition in night-time environments, researchers have used images captured with sensors that operate outside the visible portion of the electromagnetic spectrum, most often in the infrared spectrum. Recognition using face images from different spectral bands is commonly referred to as cross-spectral face recognition (CFR). CFR is more challenging than traditional visible spectrum FR due to three factors: (1) large intra-spectral variation, when in the same modality, face samples of the same subject can display larger appearance variations than face samples of different subjects, (2) the modality gap, when appearance variation between two face samples of the same subject can be larger than two samples belonging to two different subjects, and (3) a limited availability of training data of cross-modality face image pairs can make

it difficult to design successful CFR methods using deep neural networks Anghelone et al., 2022.

Earlier methods for cross-spectral face recognition with MWIR images used CNNs to learn a shared feature representation between each spectral band, like the work in Sarfraz and Stiefelhagen, 2017. More recently, synthesis methods have become very popular C. Chen and Ross, 2019; Iranmanesh et al., 2020; Peri et al., 2021a; T. Zhang et al., 2018. The goal of synthesis methods is to learn a mapping between the thermal and visible domains. In this work, I transform thermal images into the visible domain so that features can be extracted from the transformed images.

Following recent trends, I also use a synthesis technique to perform CFR so that retraining of a FR model is not necessary. Training face recognition models requires large amounts of data and time, and by choosing to use a synthesis model, I use the limited available data in a more efficient manner. I also incorporate soft biometrics into the synthesis pipeline. Soft biometric traits are physical, behavioral, or adhered human characteristics, classifiable in predefined human compliant categories Dantcheva et al., 2011. They include, but are not limited to age, race, gender, hair, scars, and tattoos. The have been used to improve recognition performance by fusing scores between the primary and soft biometric, and also by filtering the search space of the gallery database.

In this work, I use soft biometrics to improve thermal-to-visible face synthesis, and therefore, face recognition performance. Several methods investigate face recognition and gender classification in the different IR bands using synthesis, but to the best of my knowledge, no other works attempt to classify race using MWIR in the open literature in any capacity.

1.5 Other Applications

1.5.1 Firearm Detection

In this section I propose a method for firearm detection using images from surveillance footage for the purpose of quickly alerting authorities to active shooter incidents. The main contributions of this work are:

- Designing a Faster R-CNN firearm detector for both handguns and long guns specifically for surveillance video.
- Generate a novel small arms database composed of only real surveillance footage, which was manually annotated to establish a baseline.

- Determine the combination of data augmentation techniques that boost firearm detection performance.

The quick and accurate detection of firearms in surveillance and CCTV systems can aid law enforcement and first responders in preventing loss of life in many violent situations. According to the Federal Bureau of Investigation (FBI) Uniform Crime Report, there were 10,265 homicides committed with firearms in the United States in 2018. While many of these homicides are not captured on camera, active shooter incidents, which the FBI defines as one or more individuals actively engaged in killing or attempting to kill people in a populated area, often take place in public areas where surveillance systems are in use. The FBI designated 27 shootings as active shooter incidents in 2018 FBI, 1999. These incidents often resulted in tremendous consequences, including the loss of human lives. Thus, it is paramount to respond quickly to these incidents by alerting the police and other law enforcement authorities on time. While different strategies can be used to achieve such a goal, a potential solution is the use of automated firearm detection algorithms that operate real-time on security video camera footage in which such incidents may occur. These detections must then be confirmed by a human operator. However, the detection of firearms in surveillance video is very challenging. As identified in Tiwari and Verma, 2015a, firearms are very difficult to detect in operational scenarios due to the weapons being partially occluded, as well as variations in firearm shapes, camera angles, firearm pose, noise, illumination, and scale. There are also unique and diverse backgrounds in these videos captured by different camera sensors. Finally, false positives must remain low so that the human operator, whose task is to confirm detections, will not ignore the detection system because false detections are far too common. In order to identify firearms quickly with a focus on minimizing false positives, a large and diverse training set is needed to support an advanced deep learning object detection model.

In this work, I propose a Faster R-CNN Ren et al., 2015 firearm detection model that uses a ResNet-50 K. He et al., 2016 base network previously trained on the COCO Lin et al., 2014 dataset. I trained this model on a novel database consisting of only real-world images from surveillance and CCTV video where firearms are present. In order to supplement the size of the database, I performed a comprehensive assessment of data augmentation techniques in order to identify the most efficient combination in terms of firearm detection performance. Hence, I comprehensively assess eighteen (18) different data augmentation techniques while training the models in order to identify the ones that yield the highest detection performance.

My experiments show that a Faster R-CNN-based model, trained solely on challenging surveillance footage, can achieve high detection accuracy for handguns and long guns. Specifically, the proposed model can accurately detect firearms in video frames taken from real surveillance footage close to real-time, yielding precision and recall scores of 93.9% and 96.4% for handguns, and 95.2% and 94.6% for long guns. To the best of my knowledge, this is the first time in the open literature where an object detection model was trained solely on real surveillance footage for both handguns and long guns.

Chapter 2

Literature Review

2.1 Acquisition and Preprocessing

2.1.1 Facial Attribute Analysis: Mugshot Data

Forensic biometric systems are available for many modalities including face, sketch-to-photo-face, fingerprint, ear, forensic speaker recognition, and soft biometrics like scars, marks, and tattoos (SMT). In situations where primary biometrics like face and fingerprints are not available or sufficient, tattoos that are often collected by law enforcement to aid in identification are commonly used. In Lee et al., 2012, the authors proposed the Tattoo-ID automatic tattoo matching and retrieval system, which extracts SIFT keypoints and then uses a matching algorithm to measure visual similarities between the probe and gallery images before retrieving the database images with the largest similarity. It proved to be a significant improvement over using the ANSI/NIST-ITL1-2011 standard that uses defined classes to query tattoo images Wing, 2013.

One of the most valuable tools for forensic biometrics is face recognition Bourlai, 2016; Bourlai, Narang, et al., 2012. Biometric face recognition can aid law enforcement in several ways, including the detection of multiple records in a database, an additional method of identification when fingerprints or other information may not be available, rapid identity checks in the field, and a lead generator for investigations. Perhaps most importantly, biometric FR systems can quickly return a list of potential suspects to forensic operators who must manually perform the final identification of a suspect, leading to improved efficiency in both time and recognition accuracy. Returning accurate candidate lists is especially important due to the inherent human error when conducting face recognition. In White et al., 2015, the authors tested human performance on FR candidate lists of adults and children. Results showed very poor face

matching performance, with untrained participants making over 50% identification errors and trained participants making 20% fewer errors. Often, face recognition scenarios require the investigator to match low quality images captured in uncontrolled conditions against a very large database. Many of the face recognition and image retrieval challenges in forensics are discussed in Jain et al., 2011 and Jain et al., 2012. Some of the major challenges in unconstrained face recognition are variations in pose, expression, occlusion, age, and image quality factors such as illumination, blurriness, and brightness. To improve face recognition performance it is important to identify which images in a database have these attributes so that they may be further analyzed or enhanced. The three factors I focus on are (1) whether a subject's eyes are open or closed, (2) whether the subject is wearing glasses or not, and (3) whether the facial pose of the subject is either frontal or non-frontal.

- **Eyes are Open or Closed:** Detecting the eyes in face images is an important step in many automated face recognition algorithms and facial landmark localization Jain and Li, 2011. Much like face detection, the eyes have variations in appearance due to size, pose, rotation, occlusion from glasses, opening and closure of eyes, and illumination conditions X. Ding and Wang, 2011. Common factors such as closed eyes and glasses can affect different eye localization methods as observed in Bourlai and Jafri, 2011; El-Sayed and Khafagy, 2014; Whitelam and Bourlai, 2015 and therefore, FR systems. Several studies have shown that face normalization schemes based on the centers of the eyes contribute to decreased face recognition performance if eye locations are inaccurate Dutta et al., 2015 or eyeglasses are occluding the face Du and Su, 2005. To overcome some of these challenges, for example, law enforcement has used image editing and enhancement techniques of probe images, such as manually replacing closed eyes with open eyes to yield additional and more accurate returns, leading to thousands of arrests M. Taylor, 2017.

  The classification of open and closed eyes has applications in various fields including driver drowsiness detection, facial expression classification, and iris recognition. Extensive research has been done in this area using various methods including feature based González-Ortega et al., 2013; Ji et al., 2018; Kim et al., 2017; Mandal et al., 2017; L. Zhao et al., 2017, motion based Fogelton and Benesova, 2016; Hassan et al., 2016; Radlak and Smolka, 2012, and appearance-based techniques Y. Dong et al., 2016; Eddine et al., 2018; Pauly and Sankar, 2015. More recently, Ji et al., 2018 detected eye state by extracting contour features that are fitted by extracting sclera border points before determining eye state using a proposed

eyelid closure value. In Kim et al., 2017, a deep residual Convolutional Neural Network (CNN) structure was trained and tested with images collected in two different environments, achieving a lower equal error rate (EER) for classification when compared to other CNN methods like AlexNet, GoogLeNet, and other non-CNN based methods. The work from L. Zhao et al., 2017 combined a deep neural network and a deep CNN to construct a deep integrated neural network for characterizing useful information in the eye region using a joint optimization method and a transfer learning strategy to extract effective abstract eye features and improve classification capability in uncontrolled scenarios. Their experiments showed that the proposed method outperformed current state-of-the-art methods.

- **Wearing Glasses or Not:** Eyeglasses are the most common occurrence of facial occlusions and have a significant effect on face recognition performance. Not only do eyeglasses occlude the face, eyeglass frames can also be used to intentionally fool FR systems like the frames proposed by Sharif et al., 2016. The quick and accurate detection and, if necessary, removal of eyeglasses can be a critical factor in forensic biometric scenarios.

  The detection and removal of eyeglasses has been thoroughly studied and methods fall into two main categories, conventional handcrafted features Alorf and Abbott, 2017; Lazarus and Gupta, 2016; Mohammad et al., 2017; Ying et al., 2014; M. Zhao et al., 2018 and deep learning approaches Basbrain et al., 2017; LIANG et al., 2017; Y. Wang et al., 2018. In Ying et al., 2014, filtered edge intensities on grayscale images were used to determine the presence of glasses before using PCA reconstruction and inpainting to extract and remove the glasses respectively. Alorf and Abbott, 2017 used local descriptors and support vector machines to detect eye state, mouth state, and presence of glasses to achieve state-of-the-art performance when compared to CNN methods. In M. Zhao et al., 2018, a method for eyeglasses detection, location, and a frame discriminant based on edge information was proposed. By finding the horizontal and vertical nose bridge, the existence of eyeglasses is determined and the location was found using a bidirectional edge information projection. The authors then checked the existence of frames and can measure frame width based on the location of the left and right glasses. An eyeglasses detection framework based on a shallow CNN was created in Basbrain et al., 2017. Using the pre-trained GoogLeNet architecture fine-tuned for images with and without eyeglasses, the learned weights from

GoogLeNet are copied to the corresponding layers in the shallow CNN and used as a feature extractor to be classified by a trained linear SVM. The shallow architecture CNN reduced detection time by almost a factor of two while retaining high detection accuracy. Y. Wang et al., 2018 proposed a facial obstructions removal scheme based on an Enhanced Cycle-Consistent Generative Adversarial Network (ECGAN) for face recognition. Eyeglasses were used as facial obstructions, which were detected using a CNN. The eyeglasses are then removed using the ECGAN, improving accuracy of face recognition compared to other existing approaches.

- **Pose is Frontal or Non-Frontal:** Face recognition with non-frontal pose is another common problem that has yet to be completely solved and degrades FR performance C. Ding et al., 2015. The same is true of face recognition with frontal pose, where changes in terms of roll, pitch, and yaw angles also impact FR performance. Examples of several techniques to handle face recognition across pose are discussed in W. Deng et al., 2017; C. Ding et al., 2015; Ho and Chellappa, 2013; Masi et al., 2018; Oh et al., 2018; Shao et al., 2018; Xu et al., 2018; X. Zhang and Gao, 2009; X. Zhang et al., 2015. In Ho and Chellappa, 2013, pose variations are handled by a method for reconstructing the virtual frontal view from a given non-frontal face image using Markov random fields and a variant of the belief propagation algorithm. The approach divides the input image into overlapping patches, estimating a globally optimal set of local warps to transform the patches to the frontal view. Oh et al., 2018 proposed an analytic Gabor feedforward network to handle pose invariance. The network works directly on raw face images using a single sample per identity, and produces directionally projected Gabor magnitude features in the hidden layer. Next, several sets of magnitude features obtained from various orientations and scales are fused in the output layer for classification. The work in Masi et al., 2018 handled extreme out-of-plane pose variations. Using their proposed Pose-Aware Models (PAM), face images were processed using several pose-specific deep CNNs. 3D rendering synthesized multiple face poses from input images to train the models and provide additional robustness to pose variations at test time. Their results showed the approach outperformed existing methods evaluated on the IARPA Janus Benchmarks A (IJB-A) and PIPA datasets.

2.1.2 Facial Attribute Analysis: Cellphone Data

According to Zheng et al., 2018, Facial Attribute Estimation can be generalized into two groups, part-based methods K. He et al., 2018; Kalayeh et al., 2017; Z. Liu et al., 2015b; N. Zhang et al., 2014 and holistic methods Cao et al., 2018; Q. Dong et al., 2017; C. Huang et al., 2016; Lu et al., 2017. Part-based methods locate facial attributes in the image before extracting features and making an attribute prediction. In contrast, holistic methods emphasize learning attribute relationships and estimating facial attributes without any of the extra localization modules commonly used with parts-based methods.

Part-based Methods

In N. Zhang et al., 2014, a part-based attribute classification method for people under variations of viewpoint, pose, articulation, occlusion, and appearance was presented. Employing powerful CNNs, their method used poselets to help eliminate viewpoint and pose variations so that the CNN can learn pose-normalized appearance differences. By then concatenating the features of each poselet and adding a deep representation of the entire input image, the authors created a generic feature representation that achieved state-of-the-art results on the Berkeley Attributes of People Bourdev et al., 2011 and Labeled Faces in the Wild (LFW) G. B. Huang et al., 2008 datasets.

Semantic segmentation was used to improve facial attribute prediction in Kalayeh et al., 2017. To perform predictions, the authors used a CNN that generates feature maps that are aggregated and sent to a classifier. Their model learns where to attend and how to aggregate feature map activations, funneling the attribute signals into semantic regions in an approach they call Semantic Segmentation-based Pooling (SSP). By changing the max pooling operation so that it does not mix activations in different semantic regions using a gating mechanism, they also incorporated semantic segmentation into the earlier layers of the network, named Semantic Segmentation-based Gating (SSG). They also demonstrated that semantic face parsing improves with the presence of face attributes, showing the benefits of jointly modeling these two tasks.

The work by Z. Liu et al., 2015b is another part-based method for predicting face attributes in the wild. The authors proposed a framework that cascades two CNNs that they call LNet and ANet. LNet is pre-trained on massive general object categories for face localization, and ANet is pre-trained on massive face identities for attribution prediction. This strategy showed that face localization and attribute prediction performance can be improved using different pre-training strategies for each network. As a result, their final framework was

robust to background and face variations and able to use images of arbitrary size with no normalization required. A dual-path CNN to learn facial attributes was proposed in K. He et al., 2018. A facial abstraction image containing local facial parts and texture information was created using a Generative Adversarial Network (GAN). Then, features from the abstraction image and original image were fused to learn all of the attribute tasks. Their method showed improvement over state-of-the-art methods on the CelebA Z. Liu et al., 2015a and LFWA datasets.

Holistic Methods

A holistic method for face attribute classification on large-scale imbalanced data was proposed in C. Huang et al., 2016. The authors first validated classic methods for dealing with class imbalanced data. Then, they showed that more discriminative deep representations can be learned. These representation can maintain inter-cluster and inter-class margins that are expected to effectively reduce the class imbalance that is innate in the local neighborhood data. Through extensive testing, their proposed quintuplet sampling with triple-header loss was shown to work very well for imbalanced learning. Imbalanced learning in facial and clothing attributes was investigated in Q. Dong et al., 2017. Their proposed Class Rectification Loss (CRL) model used batch-wise incremental hard positive and negative mining of the minority classes to regularize the learning behavior on large scale data with substantial imbalanced class distributions. Their end-to-end framework showed the advantages of using CRL on the CelebA facial attribute and X-Domain clothing attribute Q. Chen et al., 2015 datasets.

A CNN based, scenario-dependent, and mobile device adaptable hierarchical classification framework was proposed by Narang et al., 2017 to automatically categorize face data captured under challenging conditions, which was followed by using face recognition algorithms. A multi-sensor database using 4 different phones was collected for face images indoors and outdoors using yaw angles from -90°to +90°at two different distances, 1 and 10 meters. After performing face detection and pose estimation to classify the face images into a frontal and non-frontal class, a tri-level hierarchical classification was performed where: Level 1, the face images are classified by phone type; Level 2, face images are classified as indoor or outdoor images; and Level 3, face images are classified as close (1-meter) and far (10-meter) categories. Their results showed that the proposed data groupings, when used before face matching is performed, resulted in significantly improved rank-1 identification when compared to the original all vs. all biometric system.

In X. Yang and Bourlai, 2018, the authors addressed the problem of human identity recognition using off-angle faces. Their system is composed of a physiology-based human clustering module and an identification module that is based on facial features collected from face videos. First, they passively extracted breath as an important vital sign. Then, human subjects were clustered into nostril motion versus nostril non-motion groups, a set of facial features were localized, and finally, feature extraction and matching were performed. This approach achieved improved identification rates on all datasets used, and significantly higher identification rates with the use of a single or a combination of facial features.

An automated approach for learning the structure of multitask deep learning architectures in the context of person attribute classification was proposed in Lu et al., 2017. Starting with a thin network model, their method expands the model during training using a multi-round branching mechanism. Within each layer of the network, the tasks that share features are identified, while model complexity is penalized. The proposed method is also not dependent upon the underlying task, meaning that it can be applied to other multitask problems outside the primary scope of person attribute classification. Another holistic method was developed in Cao et al., 2018 where identity information and attribute relationships are considered simultaneously to address multitask face attribute learning. The authors proposed the Partially Shared multitask Convolutional Neural Network (PS-MCNN), in which task relation is captured by a shared network and variability across tasks is captured by task specific networks. Then, by introducing a local learning constraint, the Partially Shared multitask Convolutional Neural Network with Local Constraint (PS- MCNN-LC) network was created. This network minimizes the differences between the representations of each sample and its local geometric neighbors with the same identity, which they call Local Constraint Loss. Experimental results on the CelebA and LFWA datasets showed the effectiveness of their methods.

Methods for Active Authentication

One current area of research using facial attributes is active authentication on mobile devices Niinuma et al., 2010; Samangouei and Chellappa, 2016; Samangouei et al., 2015, 2017; Smith-Creasey et al., 2018. Actively authenticating users after an initial login ensures that intruders do not gain access to system resources until the user logs out. An early method for continuous user authentication was developed in Niinuma et al., 2010. Instead of using hard biometric traits like face to continuously authenticate a user, their framework used soft biometrics, in this case clothes and facial skin color. Each time a user logs on, the

soft biometric traits are enrolled and fused with the conventional authentication schemes, password, and face biometrics. Results showed that the system is robust to the users posture and leads to better security. Binary facial attribute classifiers trained on the PubFig Kumar et al., 2009 dataset were used for continuous authentication on smartphones in Samangouei et al., 2015. Forty four (44) different classifiers used an image of the current mobile device user's face to extract attributes. Next, authentication is performed by comparing the differences between the enrolled and acquired attributes of the user. Their experiments on unconstrained mobile face data showed the method captures important face attributes and improves verification.

In Samangouei and Chellappa, 2016, a multitask, part-based Deep Convolutional Neural Network (DCNN) for attribute detection that enables continuous authentication on a mobile device was proposed. Each network in the architecture predicts multiple face attributes from a face component by mapping to a shared embedding space. Consequently, by investigating the subspace clusters of the embedding space, new attributes are also extracted. Their architecture outperformed previous attribute-based authentication methods and was efficient in terms of speed and power consumption when used on a mobile device. The authors in Smith-Creasey et al., 2018 also used facial features for continuous face authentication on mobile devices. They introduced face tracking to prevent attacks between re-authentication and liveness detection to eliminate spoof attacks. They used a novel dataset to show that there are differences in face recognition and tracking performance when a user is performing different activities (walking, sitting, and standing).

The aforementioned studies propose the detection of facial attributes when using primarily popular publicly available face datasets that were not collected using a mobile device. In the works where the researchers used mobile phones to collect face data, the standoff distances are very close to the user being authenticated. In this work, I use a set of challenging mobile-based face datasets, using data captured from traditional cameras and mobile devices, when operating at multiple standoff distances (1 meter and 5 meters), in either indoor or outdoor conditions. I address the problem of facial attribute analysis by focusing on a set of specific facial attributes for the purpose of improving face recognition based human authentication on mobile based applications. I propose training both conventional and deep learning based classifiers on challenging mobile face data. I classify three important face attributes and yield highly accurate performance results for each classifier, independent of the face dataset used. Specifically, the proposed attribute-specific detection models are robust, yielding up to 100% accuracy (in terms of F1 score) depending on the attribute tested.

2.1.3 Facial Attribute Analysis: Masked Data

Masked Face Recognition

One of the major challenges in still image-based face recognition is the partial occlusion of the face in unconstrained environments Guo and Zhang, 2019. While there is a large amount of literature dedicated to the facial occlusion problem, there are now several publications that have been reported after the COVID-19 outbreak that focus specifically on face masks. One of the first studies was performed by the National Institute of Standards and Technology (NIST) on the performance of face recognition algorithms on masked faces Ngan et al., 2020. All algorithms used in the study were provided to NIST before the pandemic, thus offering a verification benchmark for algorithms not specifically developed to handle masked face images. The occlusions were made by synthetically applying masks of different shapes, colors, and nose coverage to the probe images. Experimental results showed that the overall accuracy using masked probe images led to a substantial performance decrease for all algorithms used, and masks that covered more of the face resulted in more false non-matches.

In Damer et al., 2020, the authors also assessed the effects of face masks using their own database designed to simulate realistic use cases of people with and without masks covering their faces. They assessed two high-performing academic algorithms and one efficient Commercial Off the Shelf (COTS) algorithm, finding that masks have a large impact on score separability between genuine and imposter comparisons in all three methods. Their dataset was extended to include more participants with real and simulated masks in Damer et al., 2021. They compared the effect of masked faces on verification performance by evaluating 12 human experts and 4 popular face recognition algorithms. Among several observations, they found that human experts and the verification performance established by the algorithms are similarly affected when comparing masked probes to unmasked references or pairs of masked faces. More recently, the Masked Face Recognition Competition was held in the International Joint Conference of Biometrics 2021 and summarized in Boutros et al., 2021. The competition included ten teams from academia and industry from nine different countries. The submissions were evaluated on a database of individuals wearing real face masks on two different scenarios, masked vs masked face verification accuracy, and masked vs non-masked face verification accuracy. Ten of the 18 solutions submitted by the teams were able to achieve lower verification error than the ArcFace baseline.

Another challenge is face recognition in the thermal band, where there can be either same spectral face matching (thermal to thermal), or cross-spectral face

matching (visible to thermal). Face recognition in the MWIR band Bourlai, 2016; Bourlai, Ross, et al., 2012; Hu et al., 2015; Mokalla and Bourlai, 2019; Osia and Bourlai, 2017a is an active area of research that can be applied to a variety of surveillance or law enforcement applications. The face recognition pipeline is often similar to that of the visible, where faces must be detected, normalized, and matched. These challenges are addressed in Mokalla and Bourlai, 2019, where face detectors were trained and assessed on thermal data captured at 5 and 10-meter distances, both indoors and outdoors. Then, same-spectral cross-scenario (indoor vs outdoor) face recognition was used to compare faces detected using the trained models versus the annotated ground truth faces.

Mask Detection and Classification

Interest in the problems of detection and classification of masked faces has increased since the start of the COVID-19 pandemic. Earlier works Ge et al., 2017; J. Wang et al., 2017 used the term "masked" as a description of faces that are occluded in some way, and not necessarily from a homemade or medical face mask. The more recent works focused on the localization and classification of medical or cloth face mask occlusions only. Of the face mask occlusion literature recently produced, most detect either the presence or absence of a mask on a human face Abbasi et al., 2021; Chowdary et al., 2020; Jiang et al., 2020; Khandelwal et al., 2020; Loey et al., 2021a, 2021b; Mohan et al., 2021; Rahman et al., 2020; Sethi et al., 2021; Suresh et al., 2021, and not whether or not it is being worn correctly.

In Jiang et al., 2020, one of the first dedicated face mask detectors, RetinaFaceMask, was proposed. The one-stage detector utilized a feature pyramid network with two novel additions to increase masked face detection. A context attention detection head focuses on detecting masks, and a cross-class object removal algorithm removes objects with low confidence and high IoU scores. The authors assessed two different backbone architectures, namely a ResNet architecture for high computational scenarios and a MobileNet architecture for low computational scenarios. Results on images from the MAFA Ge et al., 2017 and WIDER Face S. Yang et al., 2016 datasets achieved state-of-the-art results. In Rahman et al., 2020, an automated system to detect persons not wearing a face mask in a smart city network was proposed. The authors created their own novel CNN-based detection architecture that can accurately detect faces with masks. Then, the decision is forwarded to a proper authority to ensure precautionary measures are being maintained. An ensemble face mask detector that uses a one-stage and two-stage detector during preprocessing was created in Sethi et al., 2021. This approach resulted in high accuracy and low infer-

ence time. The inclusion of a bounding box transformation that improved mask localization performance allowed their model to achieve higher precision in face and mask detection when compared to the previously mentioned RetinaFaceMask detector. The authors also addressed the large class imbalance of the MAFA dataset by creating a balanced version where the imbalance ratio is nearly equal to one.

Loey et al. proposed two different face mask detection methods. In Loey et al., 2021b, classical machine learning and deep learning methods were combined for accurate mask detection on three masked datasets. ResNet50 was used for feature extraction, and SVM, decision trees, and ensemble algorithms were used for classification. The proposed model outperformed related works, with the SVM achieving 99.64% on the Real-World Masked Face Dataset Z. Wang et al., 2020. In Loey et al., 2021a, the authors used a conventional deep learning approach for face mask detection. In this work they again used the ResNet50 model for feature extraction and the YOLOv2 Redmon and Farhadi, 2017 detector. Using the ADAM optimizer and mean IoU for estimating the number of anchor boxes, they achieved better results than the related work reported in their paper.

There are other works that not only detect whether a face mask is present on the face, but if it is being worn correctly Batagelj et al., 2021; Chavda et al., 2021; Qin and Li, 2020. In Qin and Li, 2020, the SRCNet was proposed for face mask detection. The method consists of a super-resolution network and face mask condition classification network that classifies three face mask wearing conditions: no face-mask wearing, incorrect face-mask wearing, and correct face-mask wearing. SRCNet applies super-resolution to all cropped faces, when width or length are no more than 150 pixels. After the network enhances the image to an output size of $224 \times 224 \times 3$, face mask condition classification is performed. An ablation study showed that both transfer learning and the super-resolution network greatly contributed to the accuracy of SRCNet. Chavda et al., 2021 used a two-stage CNN architecture for detecting masked and unmasked faces that can be used with CCTV footage. The authors constructed their own database from several publicly available masked datasets and online sources. They noted that the dataset contains improperly worn face masks and palms masking the face, and they labeled those instances as non-masked faces. After training several face detectors and classifiers, the results yielding the highest scores were achieved using the RetinaFace face detector and NASNetMobile classifier. Results on video data were also improved with a modified Centroid Tracking technique from Nascimento et al., 1999.

One of the largest studies investigating the proper wearing of face masks was carried out in Batagelj et al., 2021. The authors investigated three important research questions, (a) how well do existing face detectors perform on masked face images, (b) is it possible to detect compliant placement of face masks, and (c) are existing face mask detection techniques useful for monitoring applications during the current pandemic. To address these questions, they performed a comprehensive examination of seven pre-trained face detection models for masked face detection performance and 15 classification models for correct face mask placement. To implement the study, the authors also created the Face-Mask-Label Dataset (FMLD), compiled from the MAFA and Wider Face datasets. Most existing techniques only detect the presence of a face mask, so the FMLD dataset is annotated for compliant and non-compliant face masks. The dataset is also made publicly available. The authors, by evaluating the face detection and classification stages separately, found that RetinaFace and ResNet152 yielded the highest performance. Their results indicated that masked faces are a challenge for most face detectors, but RetinaFace was able to achieve an average precision of 92.93% on the entire dataset. The classification models performed better, with all methods achieving an average recognition accuracy of over 97% and only a 1.12% difference between the least and most accurate models in terms of accuracy performance.

Due to the COVID-19 pandemic, research on the detection and classification of face masks has been studied with increasing urgency. However, at this time there are still relatively few publications in the area, especially works that specifically classify different levels of face mask compliance. Furthermore, at the time of this writing, I could not find any literature classifying face mask compliance in the thermal band. Due to the limited number of publications in this area, I focus on the classification of face mask compliance in both the visible and MWIR bands.

2.1.4 MultiSpectral Face Dataset Collection

In table 2.1, the MILAB-VTF(B) dataset that I took a lead role in collecting is compared with similar datasets that were created in different spectra. Specifically, the MILAB-VTF(B) raw dataset is characterized by the following key advantages compared to the other thermal-visible datasets:

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

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

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

Dataset	Modalities	Subjects	Variability	IR Resolution (W × H)	Range (m)
UND Kevin and Bowyer, 2003	LWIR, Visible	241	I, E, T	320 × 240	Unspecified
NVIE S. Wang et al., 2010	LWIR, Mono	215	I, E, G	320 × 240	0.75
ULFMT Ghiass et al., 2018	MWIR, Visible	238	P, E, T, G	640 × 512	1.0
ARL-MMFD Hu et al., 2016	P-L, LWIR, Visible	111	E	640 × 480 (LW)	2.5, 5.0, 7.5
Tufts Panetta et al., 2018	NWIR, LWIR, Visible	100	P, E	336 × 256	1.5
ARL-VTF D. Poster et al., 2021	LWIR, Visible, Mono	395	P, E, G	630 × 512	2.1
MILAB-VTF(B)	MWIR, Visible	400	P, L, I_N, E_N, O_N	1280 × 1024	1.5, 100, 200, 300, 400

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

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

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

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

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

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

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

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

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

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

2.2 Keypoint Detection

2.2.1 Facial Landmark Detection

Table 2.2: Summary of recent thermal facial landmark detection approaches. $^{\mp}$ denotes number of images, not subjects.

Publication	Dataset	# Subjects	# Landmarks	Imaging Sensor	Synthesis	Model
Riggan et al., 2018	ARL Volume 1 Hu et al., 2016	60	68	Polarimetric LWIR	Yes	DLIB King, 2009
Kopaczka et al., 2018	RWTH Kopaczka et al., 2018	90	68	LWIR	No	DAN
Kopaczka et al., 2019	RWTH Kopaczka et al., 2018	90	68	LWIR	No	CNN w/PCA
D. Poster et al., 2019	ARL Vol 1, 2 H. Zhang et al., 2019	111	5	Polarimetric LWIR	No	DAN, MTCNN, PBC
Chu and Liu, 2019	RWTH Kopaczka et al., 2018	90	68	LWIR	No	multitask Unet
Keong et al., 2020	Eurocom Mallat and Dugelay, 2018a	50	68	LWIR	No	DMSL
Mallat and Dugelay, 2020	Helen Le et al., 2012, LFPW Belhumeur et al., 2013	$2{,}330^{\mp}$, $1{,}035^{\mp}$	68	RGB	Yes	AAM, DAN
D. D. Poster et al., 2021	ARL Vol 1, 2 H. Zhang et al., 2019, RWTH Kopaczka et al., 2018	111, 90	5, 68	Polarimetric, LWIR	No	CNN and Transfer Learning
Peri et al., 2021b	ARL-VTF D. Poster et al., 2021, MILAB-VTF(B) Peri et al., 2021a	395, 400	6, 21	LWIR, MWIR	Yes	Thermal-to-Visible Synthesis
Kuzdeuov et al., 2022	SF-TL54 Kuzdeuov et al., 2022	142	54	LWIR	No	Regression trees and U-net
My Method	ARL-VTF D. Poster et al., 2021, MILAB-VTF(B) Peri et al., 2021a	395, 400	6, 21	LWIR, MWIR	Yes	HRNet, CUT, StarGAN v2

Landmark Detection on Thermal Data

A summary of relevant thermal facial landmark approaches is presented in table 2.2. Many current methods predict landmarks directly on thermal data, while more recent approaches make use of generative models to synthesize data from the thermal to the visible domain. After synthesis, pre-trained visible models predict the landmarks. One of the first works for facial landmark detection in

the thermal band was Kopaczka et al., 2016, a face tracking method based on active appearance models (AAM) that used long wave infrared (LWIR) images. The authors assessed several methods for AAM generation and fitting using still images and video sequences from a small manually annotated thermal database. They found that a combination of DSIFT for modeling and SIC for fitting produces robust and stable results for face tracking in the thermal domain. The need for datasets containing high resolution and high-quality annotations was addressed in Kopaczka et al., 2018. The authors introduced a high-resolution thermal face database with manual annotations for 68 facial landmarks. They assessed the suitability of their data for deep learning using the Deep Alignment Network (DAN) Kowalski et al., 2017. They also assessed the landmark detection performance of active appearance models, finding that DAN outperformed their AAM-based approaches. Through this evaluation the authors concluded that when there are sufficient amounts of data with quality annotations, learning-based algorithms outperform algorithm-based approaches.

Chu and Liu, 2019 developed a network that jointly performed facial landmark detection and emotion recognition tasks on thermal images using a two-stage training mechanism and U-Net. The first stage finds the optimum parameters for U-Net, while in the second stage, landmark loss and emotion loss are minimized for landmark detection and emotion recognition, respectively. Their results showed that this multitask approach performed better than a single-task approach and was more robust for faces with different emotions. An efficient landmark detection network using CNNs and PCA was proposed in Kopaczka et al., 2019. The authors used PCA of landmark positions for generative modeling of facial landmarks. They avoided using iterative optimization in the neural network by including the PCA with a novel layer. Their network predicts model parameters in a single forward pass to achieve detection on hundreds of frames per second. They evaluated their method on visible data using the 300W Sagonas et al., 2013 dataset and on thermal data with the RWTH Kopaczka et al., 2018 dataset.

In D. Poster et al., 2019 the strengths and weaknesses of three modern landmark detection algorithms developed for visible images were assessed in the thermal domain. The authors found the cascaded shape regression method used in DAN to be the most efficient for adapting to thermal images. Furthermore, they found that even small errors during the alignment process can have a disproportionately negative impact on thermal-to-visible face verification results when compared to manually aligned images. Keong et al., 2020 proposed the Deep Multi-Spectral Learning (DMSL) network for facial landmark detection. The network contains two sub-modules, the first performs face boundary detec-

tion, and the second handles landmark coordinate detection. This architecture enables landmark detection on both visible and thermal images and when faces are partially occluded or off-pose. A method for visible-to-thermal parameter transfer learning using a coupled convolutional network architecture was presented in D. D. Poster et al., 2021. Instead of training exclusively on thermal images or requiring visible and thermal images at test time, their method allows for the use of vast amounts of available visible data for training, while only using thermal data during testing. Of the four types of parameter transfer learning methods presented, three outperform the baseline single stage version of DAN and an AAM that had been trained solely on thermal face image data.

Landmark Detection via Synthesis

Riggan et al., 2018 proposed a method for thermal-to-visible synthesis of face images for cross-spectrum verification and facial landmark detection. Their method was optimized using both global and local facial features to produce more discriminative faces than previous synthesis methods. Mallat and Dugelay, 2020 addressed the lack of available thermal face datasets with ground truth landmark annotations by performing visible-to-thermal synthesis on existing face databases. The synthesized thermal datasets then shared the same landmark annotations as the visible ones. Synthesis was performed using cascaded refinement networks trained with contextual loss on the Visible and Thermal Paired Face Database Mallat and Dugelay, 2018b. Active appearance models and DAN were trained on the synthesized data and then evaluated on low quality and high quality real thermal data. In Z. Wang et al., 2018, landmark detection was leveraged to aid in the generation of visible face images. During training, a detector extracts face landmarks from the generated images and propagates the loss of the predicted and ground truth landmarks back through the generative network. This method helps preserve the identity features of the face and also helps in generating more photo-realistic faces.

While the aforementioned approaches aim to successfully solve the problem of facial landmark detection, they are limited by either a small number of thermal face samples in the dataset or the requirement of paired visible-to-thermal face image data for the method to work. Additionally, to the best of my knowledge, no current works investigate both transfer learning and synthesis methods on the largest thermal face datasets, ARL-VTF and MILAB-VTF(B). In this work I propose a new methodology where I address these limitations and achieve competitive and, in some cases, better results in terms of the Normalized Root Mean Square Error (NRMSE). The synthesis method works without paired face images, and the two chosen datasets are at present the largest of their

kind reported in the open literature. This enabled me to use transfer learning on either one, or both, datasets and achieve low detection errors.

2.3 Feature Extraction and Matching

2.3.1 Race and Gender Classification for Cross-spectral Face Recognition

Soft biometrics are attributes that are not unique to an individual, but can be used in combination with a primary biometric trait to improve or expedite recognition performance Dantcheva et al., 2015. Two commonly used soft biometric traits used for improving face recognition are race and gender. These traits are most often used in the visible spectrum, but several works exist in the different regions of the IR spectrum. In this section I review relevant literature for race and gender classification and cross-spectral face recognition in the IR spectrums.

One of the early works to explore gender classification used NIR and thermal face images C. Chen and Ross, 2011. The authors evaluated several gender classification methods and found that SVM with local binary pattern histogram features resulted in the best performance. Their method also outperformed human subjects who classified faces in the thermal spectrum. A hybrid method for fusing visible and thermal IR images for gender recognition was proposed by S. Wang et al., 2016. Utilizing explicit and implicit fusion techniques, their results showed better performance for gender recognition than using only one modality. They also found that the statistical thermal features of the cheek and forehead are more reliable sources than other facial regions. A deep learning approach based on ResNet was presented by Jalil and Reda, 2022. Their proposed models achieved accuracy scores ranging from 96% to 99% and was more accurate for males than females on two thermal face datasets.

Multispectral data across the visible and NIR bands was used for gender classification by Raghavendra et al., 2018. The proposed approach was based on the Spectral Angle Mapper to extract characteristic spectral features from the high dimensional spectral data. A high classification accuracy of 93.51% was observed using their method on a dataset of 78,300 images. Gender and ethnicity classification for improving cross-spectral face recognition was investigated by Narang and Bourlai, 2016. A VGG architecture was used to classify visible and multi-distance NIR faces for Asian, Caucasian, Male, and Female groups. They found that gender classification is more accurate than ethnicity classification, and that utilizing ethnicity and gender soft biometrics resulted in significantly

improved rank-1 identification rates for cross-distance and cross-spectral scenarios.

Several synthesis methods exist in the literature for cross-spectral face recognition. Iranmanesh and Nasrabadi, 2019 presented an attribute-guided deep coupled learning framework for matching polarimetric thermal faces to visible faces. Their framework used facial attributes and multiple loss functions to learn discriminative features in a common embedding subspace. Experiments showed that their model was superior to other state-of-the-art methods on the polarimetric dataset. A method for thermal to visible synthesis by leveraging global and local regions of the face was proposed by Riggan et al., 2018. This approach provided additional regularization terms from each local region and led to better quality synthesized faces, improving cross-spectrum verification rates over other approaches and also improving facial landmark detection results. Another work using polarimetric face images was proposed by Di et al., 2019. This method used a self-attention guided GAN for synthesizing visible faces from thermal faces and thermal faces from visible. Features are then extracted from the original and synthesized images and fused for face verification.

The approach from Di et al., 2021 extracted attributes from visible images to synthesize attribute-preserved visible images from thermal images to perform cross-spectral matching. An attribute predictor network was used to extract the visible attributes, and a multi-scale generator synthesized visible images from thermal images using the extracted attributes. The proposed method achieved state-of-the-art performance on three thermal face datasets over other methods. Peri et al., 2021a studied the impact of face alignment, pixel-level correspondence, and identity classification with label smoothing for face synthesis and verification. By aligning the faces before training the synthesis model and enforcing pixel-level correspondence and feature-level identity classification during training, state-of-the-art results were observed on the ARL-VTF and TUFTS datasets.

2.4 Other Applications

2.4.1 Firearm Detection

Detecting firearms in images and video has been studied for both concealed and non-concealed scenarios. Concealed firearm detection works mostly focus on using X-ray or millimeter wave images to identify weapons. The focus of this work is the detection of firearms in the visible spectrum. Firearm detection is performed using two machine learning-based object detection techniques. Clas-

sical approaches, which use methods such as Histogram of Oriented Gradients (HOG) and Scale-Invariant Feature Transform (SIFT) for feature extraction, in combination with classifiers such as Support Vector Machines (SVM), and deep learning approaches, where neural network-based techniques are used. I review works using both techniques, with a focus on those with applications in surveillance footage. I also present a collection of research found in the open literature for firearm detection in table 2.3. In table 2.3, the "Surveillance Database" heading indicates whether the data used is composed of only surveillance footage or a combination of surveillance and other sources including TV, movies, and homemade videos. The "Surveillance Source" column indicates if the database containing surveillance footage is from real sources, created by the author, or a combination of each (if applicable).

Table 2.3: An overview of data, firearm type, and detection method of relevant works in the literature compared to my approach. It can be seen that there is a gap in the open literature for work using a large surveillance database composed of only real-world samples for the detection of handguns and long guns that has not been addressed.

Authors	Surveillance Database	Surveillance Source	Classes	Detection Method
Olmos et al., 2018	Partial	Real	Handgun	Faster R-CNN
Egiazarov et al., 2020	No	-	Long Gun	Ensemble of Semantic Neural Nets
Tiwari and Verma, 2015a	No	-	Handgun	Color Segmentation\Harris points
Romero and Salamea, 2019	Partial	Real	Handgun	VGG & ZF Net
Lim et al., 2019	Partial	Real & Generated	Handgun	M2Det
Iqbal et al., 2019	Partial	Real	Both	Orientation Aware Detector with VGG-16
Halima and Hosam, 2016	No	-	Both	Bag of Visual Words & RANSAC
Grega et al., 2016	Yes	Author Created	Handgun	Sliding Window & Neural Net
Gelana and Yadav, 2019	Yes	Author Created	Handgun	Sliding Window & Neural Net
Fernandez-Carrobles et al., 2019	Partial	Real	Both	Faster R-CNN
Proposed Approach	**Yes**	**Real**	**Both**	Faster R-CNN with ResNet-50

The intended use for firearm detection algorithms is often the monitoring of surveillance and CCTV footage. One of the early publications for pistol detection in CCTV footage used a neural network and MPEG-7-based descriptor to classify frames from a set of CCTV test movies prepared by Grega et al., 2013. In Grega et al., 2016, they again use the MPEG-7 feature descriptor to detect both firearms and knives in CCTV images. In Tiwari and Verma, 2015a a Harris corner detector with FREAK-features was used to locate firearms in images that had been color segmented using the k-mean clustering algorithm. The authors extended their work in Tiwari and Verma, 2015b, this time using SURF features. Verma and Dhillon, 2017 used a VGG-16 CNN architecture for feature extraction and SVM, K-Nearest Neighbor (K-NN), and ensemble tree classifiers, achieving over 92% accuracy with the SVM classifier. Handguns were detected in Singleton et al., 2018 by re-training a MobileNet network and using Shi Tomasi key point detection, an enhancement to Harris corner detec-

tion, to identify regions in the image for binary classification. The key point locations are used to crop out sections of the image to determine if the crop contains a firearm or not. The work in Lai and Maples, 2017 used a Tensorflow-based implementation of Overfeat-3, an integrated framework that uses CNNs for classification, localization, and detection to achieve 93% training and 89% test accuracy respectively, on images from movie, homemade, and surveillance videos.

More recently, Olmos et al., 2018 proposed an automatic handgun detection system in videos for surveillance and control objectives, with a focus on real time detection and minimizing false positives. They assessed both sliding window and region proposal classification approaches using a VGG-16-based classifier. The authors trained models on their own dataset of 3,000 firearms created with a variety of online sources. They found that a region proposal-based approach using a Faster R-CNN model obtained the highest performance with zero false positives and 100% recall on their constructed dataset and promising results when evaluated on low quality YouTube videos. In Iqbal et al., 2019, a two-phase orientation aware object detector for firearms was proposed. Phase 1 predicts the orientation of the object that is used to rotate object proposals. In phase 2, maximum area rectangles are cropped from the rotated proposals which are then localized and classified. Their method showed improved detection performance over other models like YOLO and SSD. Another two-part firearm detection system was developed in Romero and Salamea, 2019. The front end used a YOLO object detection and localization scheme to identify segments of an image where there are people. These segments were used as an input to the back end of the system, where VGG and ZF Net architectures are assessed to handle the final weapon detection. By first detecting persons in each image, only the most important areas of the image are analyzed since a weapon is most likely in the hands of a person. The authors were able to significantly reduce the number of false positives with their method. Finally, a weapon detection system is detailed in Egiazarov et al., 2020 that uses an ensemble of semantic CNN's that split up the process of detecting and locating weapons into a set of smaller problems by detecting individual weapon components. By using a dataset composed of only AR-15 rifles and identifying 4 distinct components of the rifle, their ensemble of simple neural networks was trained only to detect those specific components of the rifle. By aggregating the outputs of these networks, their proposed model showed preliminary but promising results.

2.4.2 Data Augmentation

Deep CNNs are extremely useful for most computer vision tasks. Unfortunately, to train these deep networks a large dataset is often necessary to avoid overfitting. While techniques like transfer learning are effective in helping to avoiding overfitting, it is often not enough. This is especially true when the application has no large database with which to train on. Firearm detection in surveillance video is currently one such application, so I implement data augmentation techniques to aid in the training process. Data augmentation is a solution to the problem of limited data that will enhance the size and quality of training datasets Shorten and Khoshgoftaar, 2019. I review some of the relevant works in the literature here.

Some of the most common data augmentations are basic image manipulations, namely geometric and photometric transformations. Geometric transformations include crops, flips, rotations, and translations. Photometric transformations include changes to the image RGB channels such as contrast, color, and brightness. L. Taylor and Nitschke, 2017 compared geometric and photometric transformations. They evaluated several methods on the Caltech101 dataset using a four-fold cross-validation and found that classification performance increased when applying all data augmentation methods tested.

One of the newest methods for data augmentation is generative modeling, where trained networks create artificial instances from a dataset. In Radford et al., 2015, the DCGAN was one of the early architectures to use CNNs for discriminator and generator networks. Using this architecture, the authors in Frid-Adar et al., 2018 generate computed tomography (CT) images from a small dataset of 182 images for liver lesion classification. They found that the classification performance when using only classic data augmentations resulted in 78.6% sensitivity and 88.4% specificity. When the synthetic data was added, the results increased to 85.7% sensitivity and 92.4% specificity.

While the use of data augmentation techniques can often be beneficial, researchers must be aware of application safety and other challenges when training networks with data augmentations. The safety of a data augmentation method refers to its likelihood of preserving the label post-transformation Shorten and Khoshgoftaar, 2019. This means, for example, that horizontally flipping an image for recognition of objects, such as animals or cars, will preserve the label post-transformation. But flipping a digit, such as the number 3, will not preserve the label and likely lead to problems when training the model. Additionally, Shijie et al., 2017 investigated various augmentations and combinations of augmentations on subsets of the CIFAR10 and ImageNet datasets and found that certain individual augmentations and combinations will increase performance,

while certain other combinations can degrade performance. Since the use of augmentation techniques is often domain specific, when training my detection models, it is beneficial to asses all techniques individually and in combination to identify those that will increase and decrease detection performance.

Chapter 3

Methodology

3.1 Acquisition and Preprocessing

3.1.1 Facial Attribute Analysis: Mugshot Data

Preprocessing of Faces

In order to account for the variation in image sizes between each database, I first detected faces using the MTCNN face detector K. Zhang et al., 2016 and normalized each face to 130×130 pixels. The entire cropped face image was used for classification of the frontal face factor, while the eye pairs from each face were found with a cascade object detector using the Viola-Jones algorithm Viola and Jones, 2001. The eye pairs were then cropped to 90×30 pixels to classify the eyes and glasses factors.

Feature Extraction

In this work I tested two common global feature descriptors, Histogram of Oriented Gradients (HOG) Dalal and Triggs, 2005, and Local Binary Patterns (LBP) Ojala et al., 1996. The LBP operator is a texture descriptor that computes patterns in an image by thresholding local neighborhoods, commonly 3 x 3, around every pixel in an image at the central pixel. The resulting possible 256 8-bit patterns are then converted to decimal form. The binary pattern for the pixels in a 3 x 3 neighborhood are computed as follows,

$$LBP(X_c, Y_c) = \sum_{n=o}^{7} h(g_n - g_c)2^n \tag{3.1}$$

where (X_c,Y_c) is the location of the center pixel c, n is the number of neighbor pixels, g_n is the grayscale value at pixel n, g_c is the grayscale value at c, and $h(g_n\text{-}g_c)$ is 1 if $h(g_n\text{-}g_c) \geq 0$ and 0 otherwise.

HOG features were introduced in Dalal and Triggs, 2005 for human detection and have been used successfully in a number of applications in object detection and classification. HOG features divide the image into small regions called cells, where a histogram of gradient directions are computed. To make the descriptor more invariant to illumination changes, the histograms are then normalized by accumulating a measure of local histogram energy over larger spatial regions called blocks, the results of which are used to normalize all cells in the block. The combination of all normalized histograms create the final HOG descriptor.

After several comparisons using both methods I found that HOG features consistently outperformed LBP for every factor, especially on more challenging data. Therefore, in all experiments HOG features were used for classification using a cell size of 8×8 pixels and a 2×2 block size for the eyes and glasses factors. This created a feature descriptor for each sample of length 720. The cell size for the frontal face descriptor was increased to 16×16 and used the same block size in order to reduce each sample dimensionality for training. These descriptors were of length 1764.

Conventional Models for Classification

I used 23 different models to perform the classification experiments, which included multiple Support Vector Machines Cortes and Vapnik, 1995, K-Nearest Neighbors Fix and Hodges Jr, 1951, Decision Trees Breiman, 2017, and Ensemble classifiers Dietterich, 2000. To select the best performing classification models, I performed 10-fold cross-validation on each factor in every database with all available models, creating 9 total scenarios. The results from these experiments allowed me to choose the models that generalized best to classify each of the factors across diverse data.

CNNs for Classification

In addition to using the previously mentioned models, I also trained two popular CNNs, AlexNet and GoogLeNet, on DB3 for each of the three classification factors.

AlexNet Architecture: AlexNet Krizhevsky et al., 2012 is an eight layer CNN consisting of five convolutional layers, three fully connected layers, and takes an input image of size 227×227 pixels. The output of the last fully-connected

layer is fed to a 1000-way Softmax layer which outputs probabilities for 1000 class labels. For these purposes I used transfer learning and changed the last three layers to classify 2 labels for each factor, e.g. are the eyes open or closed.

GoogLeNet Architecture: GoogLeNet Szegedy et al., 2015 is a 22 layer CNN that takes an input image of size 224×224 pixels and can also classify 1000 class labels. GoogLeNet uses nine Inception modules that convolve 1×1, 3×3, and 5×5 filters in parallel, followed by a 3×3 max pooling. I again changed the last three layers of this network to classify 2 labels for each factor.

3.1.2 Facial Attribute Analysis: Cellphone Data

In this work, I used Support Vector Machines and CNNs to classify three common face attributes using diverse databases captured with a variety of sensors.

Preprocessing

In each dataset, I accounted for image resolutions variations by, first, performing face detection using the MTCNN face detector K. Zhang et al., 2016. Then, I normalized each face to 130×130 spatial resolution. These normalized faces are used as the inputs for classifying the frontal or non-frontal face attribute. To classify the eyes and glasses attributes, eye pairs from each frontal face are detected using the Viola-Jones algorithm Viola and Jones, 2001. Next, the eye pairs were normalized to 90×30 spatial resolution and used as the inputs for the eye and glasses attributes. The number of instances where a face or eye pair are not detected by these algorithms was extremely low. In the experiments, less than 10 faces were not detected, and these samples were left out of the final database. There were almost 40 instances where an eye pair was not detected. Due to the smaller amount of frontal faces and therefore, eye pairs, these instances were manually cropped, resized, and included in the database.

Feature Extraction and Models

I used both conventional and CNN based classification algorithms. The conventional classification approaches used linear and non-linear kernels for Support Vector Machines Cortes and Vapnik, 1995, K-Nearest Neighbors Fix and Hodges Jr, 1951, Decision Trees Breiman, 2017, and Ensemble classifiers Dietterich, 2000. I also used the AlexNet and GoogleNet architectures, both pretrained on the Imagenet dataset J. Deng et al., 2009. Utilizing transfer learning, these two models were trained as binary classifiers for all three face attributes. For the conventional models, based on previous experience, Histogram of Oriented Gradients (HOG) Dalal and Triggs, 2005 features were extracted for all

experiments. Next, I performed an initial 10-fold cross-validation training on each factor in every database with all conventional classification algorithms that are available in MATLAB. This process created 9 total training scenarios. The best performing model was chosen and used in all future experiments.

3.1.3 Facial Attribute Analysis: Masked Data

Classification Models

To determine face mask compliance in the visible and thermal spectra, I assessed nine well established and pre-trained CNNs on unique dual-band face datasets. All models determine if a cropped face belongs to the "compliant" class, where a face mask is properly worn over the nose and mouth, or the "non-compliant" class, where a cropped face has either, (a) no face mask, (b) a mask worn below the nose, or (c) a mask worn below the chin. While there is no accounting for all types of improper mask wearing, especially if the mask fits snugly against the sides of the face without gaps, scenarios (b) and (c) seem to be the most common cases of masked non-compliance observed during the pandemic. All cropped faces were resized during training and inference based on the required input size of the classification model. I selected a wide array of classifiers based on model depth and number of parameters to identify any differences in performance due to model complexity on the data, see table 3.1. All networks were trained using MATLAB 2020b. The following classifiers were assessed in this work:

- AlexNet Krizhevsky et al., 2012 (2012): AlexNet heavily influenced the field of deep learning, winning the ILSVRC in 2012 by a very large margin. AlexNet features included using ReLU instead of Tanh to introduce non-linearity, and dropout regularization to handle overfitting.

- SqueezeNet Iandola et al., 2016 (2016): I use SqueezeNet v1.1 in all experiments. SqueezeNet uses 1×1 convolutions inside fire modules that squeeze and expand feature maps, reducing the number of parameters, while still maintaining accuracy.

- ResNet K. He et al., 2016 (2016): ResNet addresses the vanishing gradient problem in very deep networks by using residual blocks that allow gradients to flow through skip connections. I used three versions of ResNet: (1) ResNet18, (2) ResNet50, and (3) ResNet101.

- DenseNet-201 G. Huang et al., 2017 (2017): DenseNet is another architecture for training deeper networks. DenseNet connects every layer directly

with each other, which allows for feature reuse and reduces the number of parameters.

- DarkNet53 Redmon and Farhadi, 2018 (2018): DarkNet53 is the feature extractor used in the YOLOv3 object detector. It uses 3×3 and 1×1 convolutions and shortcut connections, which is an improvement over the previous DarkNet19 feature extractor from YOLOv2.
- NASNetMobile Zoph et al., 2018 (2018): The Neural Search Architecture (NAS) is an algorithm that learns the model architecture directly on the training dataset. NASNetMobile is a smaller version of the NASNet architecture.
- EfficientNet-B0 Tan and Le, 2019 (2019): The authors of EfficientNet use the NAS algorithm to create the baseline EfficientNet-B0 architecture, and a novel compound coefficient to scale up the network's depth, width, and resolution to improve performance.

Table 3.1: Model depth, parameters, and image input size. Parameters are in millions. *From MATLAB, the NASNet-Mobile network does not consist of a linear sequence of modules.

Model	Depth	Parameters	Input Size
AlexNet	8	61	227×227
SqueezeNet v1.1	18	1.24	227×227
ResNet18	18	11.7	224×224
RenNet50	50	25.6	224×224
DarkNet53	53	41.6	256×256
EfficientNet-B0	82	5.3	224×224
ResNet101	101	44.6	224×224
DenseNet-201	201	20	224×224
NASNetMobile	*	5.3	224×224

3.1.4 MultiSpectral Face Dataset Collection

In this section I discuss some details of the process that was followed when collecting the MILAB-VTF(B) dataset.

Figure 3.1: Overview of the data collection location and equipment setup.

Data Collection Procedures

Informed Consent and IRB Procedure To be able to de-identify the data and avoid duplicate data, the participants were assigned a random identification number associated with their biometric data after completing the registration. The participants were also asked to consent or decline the use of his/her biometric images in publications, i.e. their facial images can be used as examples in future publications that may include but are not limited to, research papers, journal articles, presentations, educational material, or other related documents.

The study involved an approximately 40-minute 2 session process, on the same day. First, an indoor session where the collection team operated using a black camping tent that was reinforced internally so that limited light was coming in. Next, an outdoor session, that involved facial image and video captured at short and long stand-off distances. The collection used state-of-the-art camera sensors from Canon (Mark IV), Nikon (PX1000) and two FLIR sensors, namely short and a long range MWIR imaging sensors (see table 3.2).

Table 3.2: Sensors used to collect the MILAB-VTF(B) dataset.

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

Data Collection

An overview of the data collection location and equipment setup can be seen in figure 3.1.

Indoor Session

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

Outdoor Session

After completion of the indoor session, the participants were taken outside to the collection area outside the tent. Participants were instructed to walk to each of the outdoor collection locations at 100, 200, 300 and 400 meters from the camera. Videos of each subject were recorded with the Nikon P1000 and FLIR 8513 cameras at each distance. Participants were instructed to turn their head as

they did in the indoor setting at each of the four different distances. Data was collected in a variety of weather conditions.

Data Structure

Training and Evaluation Protocol

The MILAB-VTF(B) dataset provides unsynchronized, paired thermal-visible videos and anonymized identifiers for each subject. Also provided are algorithmically generated frame synchronization between thermal and visible videos, face bounding boxes, and key points, which will be useful for developing end-to-end multispectral face verification pipelines.

After the collection was completed, 320 identities were selected for training and 80 identities for evaluation. Following standard face verification protocols, the collection team created gallery and query sets from the sequestered data. Specifically, four non-overlapping galleries and four non-overlapping query sets were created by splitting the evaluation data by pose (i.e., frontal/profile) and location (i.e. indoor/outdoor).

In addition to the algorithmically generated face bounding boxes and key points, there was also a manually labeled small subset of the dataset created to allow for additional evaluation. Five images from each distance, location, and spectrum for all 400 subjects are selected and labeled based on specific poses (frontal, left profile, right profile, facing up and facing down). For each image, a face bounding box and seven landmarks are annotated. The landmarks include the inside and outside corners of both eyes, the tip of the nose, and the left and right mouth corners.

3.2 Keypoint Detection

3.2.1 Facial Landmark Detection

In this section I describe the proposed approaches for facial landmark detection in the thermal domain. The first approach is landmark detection via thermal-to-visible synthesis. I start by training models to synthesize face images from the Medium Wave Infrared (MWIR) spectrum into the visible domain. Then, a pre-trained visible face landmark detector is used to predict landmarks on the synthesized faces. The performance of my method is evaluated by mapping the predicted landmarks back to the initial thermal image and is then compared with the thermal ground truth. I use StarGAN v2 and CUT for thermal-to-visible synthesis and HRNet for landmark detection on the ARL-VTF and MILAB-

VTF(B) datasets. In the second approach, I train two HRNet facial landmark detectors using visible and thermal data to learn the domain invariant features for predicting landmarks on thermal face images. The proposed method allows for the evaluation of predictions directly on the corresponding ground truth annotations. In contrast, the first method predicts landmarks from synthesized face images and maps them back to the thermal ground truth.

For all experiments in this work, I trained StarGAN v2 to perform thermal-to-visible synthesis of face images. StarGAN v2 addresses two image-to-image translation problems that many previous methods either struggle to or cannot handle: translate an image of a source domain to diverse outputs of a target domain and allow synthesis of more than two target domains. StarGAN v2 addresses these issues by generating diverse images from multiple domains. Specifically, StarGAN v2 uses two approaches for image generation, (1) a mapping network and (2) a style encoder.

- The mapping network takes a latent code as input and a domain that is sampled randomly. The output is a style code for the given domain. With multiple output branches to cover the number of domains, the mapping network learns diverse style representations for all domains.

- The style encoder takes an image and its domain as input, again using multiple output branches to account for the number of domains. The style encoder uses the input image and learns to generate a style code based on the styles from a reference image.

Using either the mapping network or the style encoder's style code and a source image as input, the generator produces an output image that reflects the style and domain of the given style code. Furthermore, because the style codes are domain specific, the generator does not need a domain label. This allows for the output of diverse images from multiple domains.

I assess both of the style code generation methods to determine which one results in the best facial landmark detection performance. The mapping network performs latent-guided synthesis where latent codes learned from the training data are sampled at random and applied to the input image. This results in a synthesized image with the same structural features as the input and one of the diverse styles learned during training. The style encoder performs reference-guided synthesis. This method applies the domain and style of a reference image onto the input image, resulting in a synthesized image with the same structural features of the input image and style of the reference image. For this work, the thermal face image is the input, and the visible face image is the reference. Additionally, I compare the StarGAN v2 results to CUT by training a model on

the ARL-VTF dataset. I compare the synthesized faces from both models by visual inspection and also in terms of facial landmark detection performance.

Thermal-to-Visible Synthesis

With the introduction of the first Generative Adversarial Network (GAN) in Goodfellow et al., 2014, image synthesis has proven to be one of the most popular applications. While traditional GANs learn to generate output images from random noise inputs, conditional GANs generate images by learning the mapping between the image input and image output. Conditional GANs, such as the pix2pix Isola et al., 2017 approach, can produce realistic results but require paired (visible - thermal) face images. The scarcity of paired data can make these methods undesirable in situations where large amounts of training data are needed. The adoption of cycle consistency loss Zhu et al., 2017 removed the requirement for paired data to perform image-to-image translation. Additionally, cycle consistency loss enforces the original reconstruction of the input image, allowing only the style of the image to change. This is very important in the context of my work, where I propose to map coordinates from a synthesized visible band face image back to its thermal input. If the source characteristics of the face cannot be preserved, e.g., the face pose, eye, nose, and mouth locations, landmark detection performance decreases.

Facial Landmark Detection using Synthesis

After the trained generative models synthesize all test set images, facial landmark detection can be performed. The ARL-VTF dataset contains six landmarks. However, I use five as in D. Poster et al., 2021 and exclude the middle mouth landmark. The five landmarks I assessed are the right and left eye centers, right and left mouth corners, and the base of the nose. I use HRNet to detect all face landmarks. The specific HRNet model was previously trained on 7,500 images from the Wider Facial Landmarks in-the-wild (WFLW) W. Wu et al., 2018 dataset and predicts 98 landmarks. I predict all 98 landmarks from HRNet and keep only the five that I require.

I established HRNet performance baselines on the ARL-VTF visible and thermal datasets for all dataset sequences; baseline, expression, pose, and glasses. The baseline sequence contains frontal images of subjects with a neutral expression. The expression sequence is also frontal face images with the subject counting out loud, starting at one. The pose sequence contains subjects slowly turning their heads from left to right. The glasses sequence is similar to the baseline and contains only subjects who naturally wear glasses. These subjects

removed their glasses for the three previously mentioned sequences. All images were captured at a distance of about 2.1 meters. After evaluating baseline HR-Net performance on visible and thermal test sets, I predicted landmarks on the synthesized face images.

Multispectral Facial Landmark Detection

An assessment of the HRNet facial landmark detector when trained on visible and thermal data is necessary to observe how well the model can generalize to both domains, as opposed to using synthesized images. To accomplish this, I fine-tuned HRNet on the MILAB-VTF(B) dataset and reported standard facial landmark detection metrics on a manually labeled subset of test data.

3.3 Feature Extraction and Matching

3.3.1 Race and Gender Classification for Cross-spectral Face Recognition

This section describes my proposed approach for cross-spectral face recognition. First, I train a multitask CNN based on the VGG architecture and an Efficient-Net B0 Tan and Le, 2019 model to classify race and gender with MWIR face images. Then, the race and gender predictions are used to select random visible face images from the MILAB-VTF(B) dataset that match the predicted race and gender. The visible images are selected from the training set since the only other option would be to select images from the test set. In order to minimize bias during synthesis, the images used for training the StarGAN v2 synthesis model and selecting the visible reference image do not overlap. Then, the selected image is used by StarGAN v2 as a reference to synthesize the MWIR face from the thermal to the visible domain. After synthesis, the MWIR face image now looks like an image captured in the visible spectrum. Face recognition is performed using ArcFace J. Deng et al., 2019 and VGG-Face Parkhi et al., 2015. A set of experiments to evaluate the effectiveness of the race and gender classification for thermal-to-visible synthesis are conducted. The experiments scenarios select the correct or incorrect race and gender before performing synthesis and face recognition. Baseline scenarios matching visible and thermal faces to visible faces are also included.

Data Normalization

All images required face detection and geometric normalization before the classification, synthesis, and face recognition steps. All faces were cropped using the ground truth bounding boxes from the MILAB-VTF(B) labels. Then, the faces were geometrically normalized using the ground truth eye centers and transformed so that the eyes lie on a horizontal plane.

Multitask Learning and EfficientNet

Two methods for classification of gender and race using MWIR face images are evaluated. The first is a network that employs multitask learning and is based on the VGG-16 architecture. Multitask learning is a method in which multiple tasks are learned simultaneously by a shared model. In this case, race and gender are the two tasks. In the multitask network, each task has it's own loss function that is optimized during training. Both tasks share early layers in the network, which can more efficiently use a small amount of data and avoid overfitting Crawshaw, 2020. Then, the network splits into separate branches for each task where the weights learned during training are not shared.

The first two convolutional layers use the pre-trained weights from the ImageNet J. Deng et al., 2009 dataset and are frozen during training. The next eight convolutional layers are shared between the race and gender tasks. The network is then split into the two tasks, each containing an additional three convolutional layers and a dropout layer set to a dropout value of 0.6. The gender task classifies images as male or female and the race task classifies White, Asian, and Black. Each task is optimized using the categorical cross-entropy.

In addition to designing a multitask network, I also use EfficientNet, a state-of-the-art classification model. EfficientNet employs a technique called compound scaling Tan and Le, 2019 to identify the optimal network depth, width, and resolution that leads to the best performance. To train the network, all layers use pre-trained weights from the ImageNet dataset and are frozen during training. A global average pooling and dropout layer are added to the top of the network to condense the output to a 1×1280 feature vector. Finally, a new softmax layer is added to output predictions for the correct number of classes. The model trained for gender classification has two outputs and the model trained for race classification has three outputs. The EfficentNet and multitask networks both use random horizontal flips as the only technique for data augmentation during training. Several other augmentation techniques were assessed but did not improve classification performance.

Reference-guided Synthesis and Face Recognition

Synthesis of the faces used was performed by StarGAN v2, using the same technique described in Section 5.2.1. The only difference in the process was that the StarGAN network was trained on the MILAB-VTF(B) dataset instead of ARL-VTF. After synthesis, the synthesized face image is matched against a gallery of visible images using ArcFace and VGG-Face. Both are state-of-the-art face recognition models that achieved 99.83% and 98.78% recognition accuracy on the Labeled Faces in the Wild benchmark dataset G. B. Huang et al., 2008. I assessed several scenarios to investigate the effect that race and gender have on face recognition when using a thermal-to-visible synthesis approach. A best case scenario where the same subject's face is used as the reference image for synthesis and for face matching is included. This is theoretically the best recognition scores that the proposed method can achieve because the correct race and gender are used to perform synthesis with the same identity. Combinations of incorrect race and gender selections are also assessed and compared against the two models that I trained to predict race and gender.

The experimental scenarios are:

- Match visible to visible faces, no synthesis.
- Match thermal to visible faces, no synthesis.
- Select the same identity for the synthesis reference image as the identity to be used for matching.
- Select a random race and gender for the reference image before synthesis.
- Select the wrong gender for the reference image before synthesis.
- Select the wrong race for the reference image before synthesis.
- Select the race and gender for the reference image using the trained EfficientNet model.
- Select the race and gender for the reference image using the trained multitask model.

3.4 Other Applications

3.4.1 Firearm Detection

In this work, I assess the performance impact of different data augmentation techniques for object detection on a novel firearm database. The following section describes the augmentation techniques and object detection model.

Database Creation

To train deep learning models that can accurately make predictions on new and unseen data, it is important to have a large and diverse training dataset. Unfortunately, there is no publicly available dataset for firearm detection in surveillance videos, so a novel dataset was created to overcome this issue. The database consists of two classes, handguns and long guns. Handguns include any type of pistol or revolver, and long guns include any type of rifle or shotgun. All database images were collected from various sources on the web including, but not limited to, videos uploaded by news stations, police departments, and self-defense and concealed carry organizations that contain footage from mostly indoor, but also some outdoor settings. After all frames were extracted from the videos, they were manually labeled with standard bounding boxes. The database consists of 11,652 frames from 90 surveillance videos.

Data Augmentation Techniques

The use of augmented data when training deep learning models is very effective in improving performance on a variety of computer vision tasks, especially object detection. The Tensorflow Object Detection API J. Huang et al., 2017 provides a wide range of augmentation methods for improving object detection performance and is used for all experiments. I chose eighteen of the provided geometric and photometric techniques for further investigation. They are applied randomly to images during training, and include crops, flips, color, and contrast changes. Table 3.3 summarizes all of the chosen methods as named in the Tensorflow API.

Table 3.3: Summary of the data augmentation techniques I tested from the Tensorflow API. A * or ** denotes a technique that improved performance in experiment 1. A ** also denotes a technique further evaluated in experiment 2

Augmentation Techniques	
Normalize Image	Horizontal Flip
Pixel Scale **	Image Scale **
Adjust Brightness **	Adjust Contrast *
Adjust Hue **	Adjust Saturation **
Distort Color *	Jitter Boxes *
Crop Image	Pad Image
Crop Pad Image	Crop to Aspect Ratio
Black Patches	Vertical Flip
Rotation 90	RGB to Gray *

Chapter 4

Experiments and Results

4.1 Acquisition and Preprocessing

4.1.1 Facial Attribute Analysis: Mugshot Data

Databases

Figure 4.1: Sample images from DB1 captured at -90°to +90°poses at 45°intervals. Additionally, every subject has an identical set of images with the eyes closed.

Figure 4.2: Sample images from DB2 with various poses, backgrounds, and illumination conditions.

- **Good Quality Face Database (DB1):** The database contains face images collected indoors at a distance of 2 meters from 1 session with a Canon

EOS 5D Mark II and Mark III camera. Images were captured from -90°to +90°poses at 45°intervals, each with the subject's eyes open and closed. Overall, the database is composed of 1719 subjects and 15240 images. This data closely represents high quality mugshot photos and is therefore used as our baseline database for classification. A sample of these images can be seen in Figure 4.1.

- **Multiple Encounter Dataset II (DB2):** This database is a collection of law enforcement submissions of deceased persons with multiple prior encounters. The dataset consists of 518 subjects collected indoors at various profile, near frontal, and frontal poses under variable illumination conditions. The sensors used to capture these images are unknown and result in a wide range of image dimensions, with 70% being approximately 0.3 mega-pixels. The number of samples per subject varies, with 262 of the subjects having 1 sample and the remaining 256 subjects having anywhere between 2 and 18 samples, totaling 1,309 images. This mugshot data represents the range of variations that can be frequently encountered in real world scenarios. A sample of these images can be seen in Figure 4.2.

- **Combined Database (DB3):** This database is the combination of DBs one and two. By combining these databases, I train classifiers that capture the variance in both high and low quality mugshot submissions.

- **Database Partitioning:** While the experiments are the same across each of the three factors, the data used for each factor is unique. The eye and glasses classification data in DB1 and DB2 are composed of only frontal face images where both eyes can be detected. In order to compensate for the low number of glasses, closed eyes, and non-frontal face samples in DB2, data augmentation is performed in order to balance the classes. Synthetic data was created to augment the eye and face factors. For the eyes, all 21 subjects with closed eyes were augmented, creating 42 additional images per subject and 882 images total. Each closed eye pair was flipped along the horizontal axis. Then these two eye pairs, the original and flipped samples, were additionally augmented with Gaussian noise, salt and pepper noise, two levels of increased contrast, two levels of increased brightness and two levels of decreased brightness, and two increasing levels of Gaussian blur. This process created 22 total images. Then, each of these 22 augmented images was given a random x and y axis translation of $\pm$ 5 pixels. For the face factor, the non-frontal face images were flipped along the x axis and Gaussian blur was added to the original, creating two additional images per sample, totaling 718 additional non-frontal face im-

ages. Lastly, the glasses factor was supplemented with subjects from the Labeled Faces in the Wild G. B. Huang et al., 2008 database, containing labeled faces that span a range of in the wild conditions including pose, lighting, race, accessories, occlusion, and background. 280 subjects with glasses and 324 subjects without glasses were used from this database to supplement DB2. A summary of the number of images by database for each factor is shown in Table 4.1.

Table 4.1: Summary of the number of images used in each scenario from every database. A * denotes the addition of augmented data.

Databases	**Number of Images**					
	Eyes Open	Eyes Closed	Glasses	No Glasses	Frontal	Non-Frontal
Good Quality Face Database	1636	1614	1174	1181	3283	6558
Multiple Encounter Dataset II	905	904*	287*	336*	940	1077*
Combined Database	2541	2518	1461	1517	4223	7635

In this work, I use CNNs, several traditional classifiers with different kernel functions, including quadratic, cubic, and Gaussian to perform classification. In the first experiments, I determine what models perform the best on the datasets using HOG features with both good quality and challenging data. I also find the models that generalize the best to the combination of those two datasets and can perform well on real world data. In the second experiment, I train and test two CNNs on DB3 to observe any improvements over using HOG features, as well as implement score level fusion of the traditional and CNN classifiers by summing the final scores from each class.

Training, Testing, and Optimization

In the experiments performed with CNNs, DB3 was split using 60% of the data for training, 20% for validation, and 20% for testing on each of the three factors. To train the networks I selected a batch size of 100 for the eyes and frontal face factors, and 50 for the glasses factor due to the smaller amount of available data. I performed empirical optimization on learning rate, epoch, and momentum parameters, repeating the same process for both networks that resulted in the best classification accuracy for each factor. First, an initial range of eight learning rates (LR) were tested, evenly spaced from 0.01 to 0.0001, holding all other parameters the same. Then a sub-range of learning rates that performed best was selected, and a final set of five evenly spaced LRs were chosen from this subset. Using each of these selected LRs, experiments for every combination of epochs from 4, 8, ..., 20, and momentum parameter of 0.6, 0.65, ..., 0.95 were conducted. An epoch value of 16 worked best for both networks in the eyes

and frontal face classifiers, and a value of 12 for both networks for the glasses. AlexNet momentum values of 0.85, 0.9, and 0.85 were the best for eyes, frontal face, and glasses respectively, and 0.95 for every experiment using GoogLeNet.

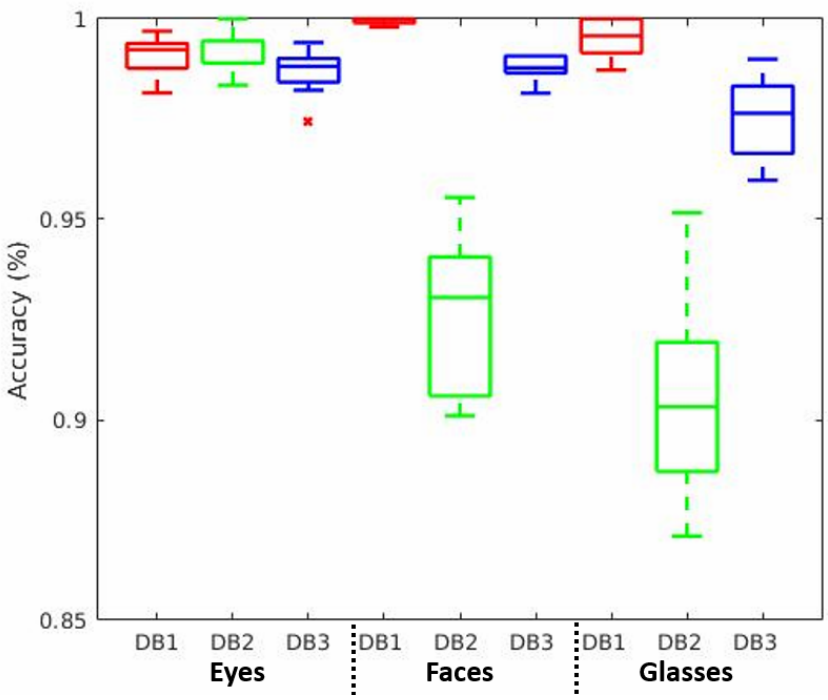


Figure 4.3: Classification results for open and closed eyes (Left), frontal and non-frontal faces (Middle), and presence or absence of glasses (Right) on the same axis.

Classification Results

For the classification of eyes, frontal faces, and glasses in Experiment 1 I found that SVMs achieved the best classification results with the exception of a Fine KNN for classifying eyes open or closed in DB2. As expected, accuracy was nearly the same or lower for DB2 in each of the three classification scenarios. This is due to the unconstrained environments in which these images were captured as well as the relatively low number of training images when compared to DB1. The classification performance across each database and each factor can be seen in the box plot in figure 4.3. It is important to note that the substantially larger number of training images in DB1 could skew the overall accuracy reported in table 4.2 on DB3 by classifying a large number of good quality images and a much smaller number of low quality images. The results in table 4.2 show the best achieved accuracy for each of the three DBs, with DB3's columns

detailing how accurate that classifier was on DB1 and DB2 data and the final accuracy on DB3. Table 4.2 also shows the accuracy of the best classifiers trained only on DB1 and only on DB2 data separately. The results show that the accuracy achieved with the classifier trained on DB3 was nearly identical to the performance when training on each dataset individually. This means that the best performing classifier of DB3 generalized very well to the combined data and can accurately classify good and poor quality images.

Table 4.2: Summary of the average classification accuracy for each factor using 10-fold cross-validation in all databases. The three DB3 columns show the average cross-validation classification accuracy in terms of the data from DB1 and DB2 individually, as well as the combined databases, DB3, in order to show how well the final trained model can classify both good and challenging data.

Average Classification Accuracy					
Factors	**DB1**	**DB2**	**DB3**		
			DB1	DB2	DB3
Eyes Open or Closed	99.0	99.2	98.7	98.5	98.7
Frontal or Non-Frontal Faces	99.9	93.5	99.8	93.2	98.8
Glasses Present or Absent	99.6	89.2	99.6	89.7	97.6

Table 4.3: Comparison of the average classification accuracy on DB3 from 10-fold cross-validation using SVMs, the best achieved accuracy from parameter optimized CNNs on DB3 test data, and fusion of SVM and CNNs.

Accuracy: SVM vs CNN					
Factors	SVM	Alexnet	GoogLeNet	SVM+AlexNet	SVM+GoogLeNet
Eyes Open or Closed	98.7	99.4	99.2	99.5	99.5
Frontal or Non-Frontal Faces	98.8	98.4	98.7	99.7	99.7
Glasses Present or Absent	97.6	99.0	99.8	99.9	99.8

In the second experiment, I optimized AlexNet and GoogLeNet to train and test on DB3 and compared classification accuracy against the SVMs from experiment 1. The results are shown in table 4.3. I observed that both CNNs improved classification of open and closed eyes by as much as 0.7% and the glasses factor by over 2%. However, the CNNs for frontal and non-frontal face classification were nearly the same as the SVM. After fusing the scores across all scenarios, I was able to achieve almost 100% accuracy for all 3 factors.

Table 4.4: Summary of the number of images used in each scenario from every database.

Databases	Number of Images					
	Eyes Open	Eyes Closed	Glasses	No Glasses	Frontal	Non-Frontal
Database 1	2,540	2,516	1,462	1,516	4,226	7,635
Database 2	471	455	606	673	1,500	1,570
Total	3,011	2,971	2,068	2,189	5,726	9,205

4.1.2 Facial Attribute Analysis: Cellphone Data

Databases and Preprocessing

- **Database 1 (DB1):** There are 2 databases used in these experiments. The first one is created from our previous work Rose and Bourlai, 2020, also reported in section 4.1.1. It consists of two separate databases that are combined to capture variations from high quality and low quality face images. The dataset with high quality face images are from a single session collected indoors at a distance of two meters using the Canon EOS 5D Mark II and Mark III cameras. I call this dataset DB1-1. Samples from each subject were collected from -90°to +90°poses at 45°intervals, with the subject's eyes open and closed for each pose.

  The second portion of this dataset, DB1-2, includes mugshot samples from law enforcement submissions of deceased persons who have multiple prior arrests. These subjects were collected indoors at various poses with different uncontrolled background and illumination conditions and unknown sensors. These uncontrolled conditions make this data very challenging to classify. In total, the database is composed of 2,237 subjects and 16,549 images. Samples from this database are shown in figure 4.4.

- **Database 2 (DB2):** The second database is collected using an iPhone 5S. Samples are captured indoors and outdoors at distances of 1m and 5m each. Data captured at 1m show the subject from the waist up, while the 5m data shows the full body of the subject. This database contains 100 subjects totaling 3928 images. Samples from this database are shown in figure 4.5.

- **Database Partitioning:** The experiments performed on all three face attributes are the same, however, the data used for each attribute is not.

Figure 4.4: Database 1 images collected from controlled (DB1-1,top) and uncontrolled (DB1-2, bottom) conditions.

The eyes and glasses attribute data are taken only from samples of frontal face images, where both eyes can be detected. As previously discussed, certain classes are imbalanced in most face attribute databases, and that is true for the databases used here. Samples where a subject's eyes are closed, glasses are present, and the face has a frontal pose are underrepresented in the data. To address this problem, random samples from DB1 and DB2 are automatically augmented to balance the classes. Several techniques, including horizontal flips, Gaussian and salt and pepper noise, random x and y axis translations, Gaussian blur, contrast, and brightness changes are used until the classes are closely balanced. The number of images in each database are summarized in table 4.4. Before augmentation, each database is split into stratified training and test sets for each of the 3 attributes, using 90% of images for training and 10% for testing. Stratifying the test set ensures that I am testing the models on data would likely be observed in real-world scenarios, i.e. open eyes will be encountered far more often than closed eyes, so I ensure that the proportion of open and closed eyes in our database is represented as such in the test set.

Experiments

To train the CNN models, I used 90% of the data for training and validation and 10% for testing. I assessed different learning rates ranging from 0.01, 0.03, 0.001,

Figure 4.5: Database 2 images from left to right, 1m indoors, 5m indoors, 1m outdoors, and 5m outdoors.

0,003, ..., 0,000001, 0.000003, while all other parameters were kept constant. Then, once I determined which learning rates are the optimal ones for the model tested, I used them with combinations of (a) epochs - ranging from 4, 8,...24, and (b) momentum values for stochastic gradient descent - ranging from 0.6, 0.65,…., 0.95 in increments of 0.05. This empirical study was conducted for both CNN based models, and results in a set of parameters yielding the highest classification performance in terms of classification accuracy.

Results

- **Baseline Classifiers on Mobile Phone Data (Experiment 1):** I classified the data in the mobile phone database, DB2, using baseline SVM and CNN models that were trained on DB1 from my previous work Rose

Table 4.5: Summary of cross-validation classification accuracy results using baseline classifiers trained on DB1 from my previous work, on DB2. SVMs that were trained on DB1, DB1-1, and DB1-2 are provided here, and also CNNs trained on DB1.

Scenarios		Indoor		Outdoor	
		1m	5m	1m	5m
Cubic SVM DB1-1	Eyes	88.94%	82.56%	67.01%	66.84%
	Glasses	81.91%	68.21%	73.60%	64.25%
	Face	98.66%	99.32%	99.83%	99.48%
Ciubic SVM DB1-2	Eyes	86.93%	77.95%	60.91%	55.44%
	Glasses	98.49%	92.31%	74.11%	61.14%
	Face	90.48%	93.57%	88.78%	88.49%
Cubic SVM DB1	Eyes	97.49%	94.87%	74.62%	73.58%
	Glasses	95.98%	90.26%	77.16%	61.14%
	Face	93.99%	94.42%	92.18%	92.44%
Alexnet DB1	Eyes	98.99%	85.13%	75.63%	61.66%
	Glasses	98.99%	93.33%	92.39%	82.90%
	Face	97.07%	83.59%	86.22%	75.95%
GoogleNet DB1	Eyes	98.99%	89.23%	76.14%	60.10%
	Glasses	99.50%	97.95%	88.32%	81.35%
	Face	94.32%	77.33%	92.18%	80.58%

and Bourlai, 2020. The face dataset scenarios in DB2 are 1 and 5 meters indoors, as well as 1 and 5 meters outdoors, captured using an iPhone 5S.

- **Retrain Classifiers on Mobile Phone Data (Experiment 2):** All four scenarios of DB2, as mentioned in the previous section, are treated as one dataset (namely they are all used as one dataset to train our models). Then, I retrain the chosen models using a 10-fold cross-validation for the SVM models tested and a 5-fold (due to computational complexity less folds were used in this study) cross-validation for CNNs.

- **Train on combined Databases (Experiment 3):** Classifiers are trained again using 5-fold cross-validation with 100% of DB1 and 80% of DB2 in every training fold, and the remainder of DB2 is always the validation set.

I hold out 10% of DB2 to use as a final test set. This test set contains no augmented data and is stratified so that the proportions of each class for each attribute are represented. Due to the number of test set samples, cross-validation

Table 4.6: Summary of cross-validation classification results from training on DB2 and training on the combined dataset of DB1 and DB2. For the combined dataset, DB2 is split into five folds, 80% is included in the training fold, and the remaining 20% is the test fold. DB1 always remains in the training fold.

SVM	**Trained on DB2**				**Trained on Combined Data**			
	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
Eyes	88.19%	90.28%	89.23%	89.10%	88.55%	89.81%	89.18%	89.10%
Glasses	99.17%	98.68%	98.92%	98.91%	100.00%	100.00%	100.00%	100.00%
Face	99.72%	99.65%	99.68%	99.68%	100.00%	100.00%	100.00%	100.00%
Alexnet	**Trained on DB2**				**Trained on Combined Data**			
	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
Eyes	86.44%	89.10%	87.75%	87.56%	86.71%	85.07%	85.89%	86.02%
Glasses	96.59%	100.00%	98.26%	98.25%	97.69%	99.66%	98.67%	98.66%
Face	99.16%	99.93%	99.54%	99.54%	98.59%	98.52%	98.56%	98.55%
GoogleNet	**Trained on DB2**				**Trained on Combined Data**			
	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy
Eyes	86.64%	92.18%	89.32%	88.98%	85.68%	87.91%	86.78%	86.61%
Glasses	95.50%	100.00%	97.70%	97.66%	97.53%	99.66%	98.58%	98.58%
Face	99.72%	99.86%	99.79%	99.79%	99.43%	98.17%	98.80%	98.80%

is always performed on the training data to account for the low statistical power of the test set.

The baseline classifiers from my previous work Rose and Bourlai, 2020 are first assessed on our mobile phone data, DB2. I report the overall accuracy using the baseline classifiers that were trained on DB1, DB1-1, and DB1-2 for all three face attributes from experiment 1 in table 4.5. Results from experiments 2 and 3 are reported in table 4.6. For all experiments, the true positive classes for each attribute are eyes open, glasses present, and frontal face, while the true negative classes are eyes closed, no glasses present, and non-frontal faces.

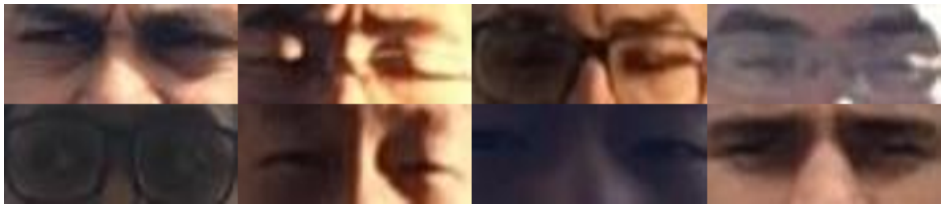


Figure 4.6: Misclassified open or closed eye images. Nearly 90% of all misclassified samples were from one of the outdoor scenarios.

- **Eye Attribute:** The best accuracy achieved for the eye attribute in experiment 1 was 98.99%. The best F1 score in experiment 2 was 89.32%, and the best F1 score achieved in experiment 3 was 89.18%. I propose using

Figure 4.7: Misclassified frontal or non-frontal face images. Almost all misclassified samples were from one of the outdoor scenarios with harsh lighting conditions or blur.

the cubic SVM trained in experiment 3 to achieve the best accuracy in terms of F1 score.

- **Glasses Attribute:** The best accuracy achieved for the glasses attribute in experiment 1 was 99.5%. The best F1 score in experiment 2 was 98.92%, and the best F1 score achieved in experiment 3 was 100%. I propose using the cubic SVM trained in experiment 3 to achieve the best accuracy in terms of F1 score.

- **Face Attribute:** The best accuracy achieved for the face attribute in experiment 1 was 99.83%. The best F1 score in experiment 2 was 99.79%, and the best F1 score achieved in experiment 3 was 100%. I propose using the cubic SVM trained in experiment 3 to achieve the best accuracy in terms of F1 score.

As I achieved a baseline accuracy of nearly 99% or better on these datasets from my previous work, see section 5.1.1, it is evident that using these classifiers on DB2 resulted in a significant decrease in performance. This is most notable with the eyes and glasses attributes, where even the scenarios that have face

Figure 4.8: Misclassified glasses or no glasses images. Almost all misclassified samples had very dark illumination where the eyes were barely visible.

images very similar to the DB1 data, indoor 1m and 5m, the classifiers performed slightly worse than the baseline. The drop in performance is more noticeable in the outdoor scenarios. There, the accuracy is nearly 30% lower when compared to the same distance indoors. Interestingly, the face classifiers seem to be quite robust to the new data by both distance and location.

For experiments 2 and 3, all trained models were able to achieve nearly 90% performance in precision, recall, F1, and accuracy for the eyes open or closed attribute, a significant improvement from the results seen in table 4.5, but still not near the 99% I achieved previously on DB1. The misclassifications for this attribute, as well as glasses and face, are almost exclusively from the outdoor scenarios where many subjects were looking into the sun. This created harsh illumination conditions and the subjects squinting their eyes so that they were barely open enough to be labeled as open.

Limitations

Samples of incorrect classifications for eye pairs, faces, and glasses are shown in figures 4.6, 4.7, and 4.8 respectively. The glasses and face attributes were able to achieve performance of 95% or better range in all reported metrics. As was observed from the first set of experiments, the face classifier continued to be robust to the distances and locations of the mobile data. The glasses classifier was also able to achieve better performance since the presence of glasses is affected less by the illumination conditions and distance changes.

4.1.3 Facial Attribute Analysis: Masked Data

Dataset

The multispectral masked database Peri et al., 2021a was collected during the winter of 2021 when mask mandates were the norm in most places. For this work, I selected 100 of the 280 subjects who participated to compose the dataset. The visible data was captured using a Canon EOS 5D Mark IV DSLR camera. It uses a 30.4MP full frame CMOS sensor and a Canon EF 70-200mm f/2.8 L-series lens. The thermal data was captured using a FLIR A8581 MWIR camera with a 50mm, f/2.5 manual focus lens. The FLIR has an indium antimonide detector and a 3.0-5.0 μm spectral range and thermal sensitivity of $\leq$ 30 mk. Examples of the visible and thermal data are visualized in figure 4.9 and figure 4.10, respectively.

Figure 4.9: Visible spectrum examples of masked and unmasked faces.

All subjects were filmed with and without a mask on, indoors, at six feet (this created a subset of our original MILAB-VTF(B) no mask face dataset). All subjects wore their mask over the nose and mouth for the masked portion of the data collection. For every subject, 10 frontal face frames were extracted from each of the videos. Then, faces were cropped in the visible band using the MTCNN face detector K. Zhang et al., 2016 and all thermal frames were

Figure 4.10: Thermal spectrum examples of masked and unmasked faces.

manually cropped. The total number of images for each spectrum can be seen in table 4.7

After the fully compliant and non-compliant faces were processed, synthetic masks were added to the compliant faces to create the two levels of non-compliant mask wearing. For visible images, the method used in Anwar and Raychowdhury, 2020 was applied. It uses facial landmarks detected by Dlib King, 2009 to identify the face tilt and six key features of the face that are required for applying the mask. These points include the nose bridge, chin, and four points along the jawline, two on each side. The 6 key points were modified to shrink and shift the mask downward to create non-compliant cases. The color code and mask type code were set to select random masks that are surgical style blue and white, and cloth style that are blue, gray, black, and dark green. For the thermal face images, for which the Dlib detector did not work, the masks were manually (via a software) applied on faces. First, a mask was extracted from a face in the visible dataset and saved as a template. Then, the mask was shrunk to a suitable size for each of the two non-compliant conditions and placed over the face. Since the thermal images do not record color information, the color of

the mask was restricted to black, similar to the masks captured with the FLIR camera. Examples of synthetic masks in the visible spectrum are presented in figure 4.11

Figure 4.11: Examples of synthetically masked faces.

Table 4.7: Compliant samples have real masks that cover the nose and mouth. Non-compliant samples have synthetically applied masks that do not cover the nose or the nose and mouth.

Spectrum	Compliant	Non-compliant	Total
Visible	1,000	1,000	2,000
Thermal	995	1,000	1,995

Experimental Setup

In this section I discuss the experiments performed on the dataset. Due to the small size of the dataset, a 5-fold cross-validation is performed. All folds are split evenly between the compliant and non-compliant classes with no overlap between samples. The different levels of non-compliant mask wearing are also evenly distributed among the folds.

For equal comparison, all models used pre-trained weights from the ImageNet database J. Deng et al., 2009 and were then trained using the same hyperparameters via transfer learning. Transfer learning is a technique to leverage the feature representations that the network has already learned and use this knowledge to learn features on a new set of images. To perform transfer learning, generally, weights in the early and middle layers, where features such as edges and textures have been learned, are frozen. Then, the final layers of the network, where more complex features have been learned, are trained to learn the new representations of the dataset. This process is especially useful when the problem to be solved does not have a large amount of data available for training.

I empirically assessed a range of learning rates and found a rate that would converge quickly during training for all models, so I use this same parameter for training every model. Optimization was performed using the ADAM Kingma and Ba, 2014 algorithm, with an initial learning rate of 0.0001 and decreased by a factor of 0.1 every 3 epochs. The models were trained for 5 epochs using a batch size of 16 on an NVIDIA GeForce RTX 2080 Ti graphics card, with 11 GB of memory. Data augmentation was performed during training using random reflection, scale, and translation changes. Training was performed 5 times for each model, once for every cross-validation fold, on the visible and thermal datasets.

Evaluation Metrics

I report the classification accuracy, which is the number of predictions the model got correct divided by the total number of samples, for all models. The formula for computing accuracy is:

$$\frac{TP + TN}{TP + FP + TN + FN} \tag{4.1}$$

where TP = True Positive, TN = True Negative, FP = False Positive, and FN = False Negative.

Results and Discussion

In this section I present the results of the experiments on face mask compliance classification in the visible and thermal spectra. The goal is to provide results of face mask compliance using established classification models in the thermal band and a complimentary visible dataset that was simultaneously captured under the same conditions.

Table 4.8: Accuracy results for the problem of thermal and visible mask compliance classification.

Model	Thermal	Visible
AlexNet	0.999	1
SqueezeNet v1.1	1	1
ResNet18	1	1
RenNet50	1	1
DarkNet53	1	1
EfficientNet-B0	1	1
ResNet101	1	0.999
DenseNet-201	1	0.999
NASNetMobile	1	0.996

Visible Results

The accuracy for all models is presented in table 4.8. Through all five folds, six of the nine models were able to classify every sample correctly, and two of the three models misclassified just a single image. NASNetMobile yielded the lowest accuracy results, missing seven samples. Five of those seven misclassifications were from the non-compliant class and the remaining two from the compliant class. The non-compliant below the nose and non-compliant below the mouth images each accounted for one of the errors, with the remainder being either no mask/non-compliant or fully masked/compliant. Some of those misclassifications are visualized on the left side of figure 4.12. Across all models, fully compliant and non-compliant with no mask samples composed most of the misclassifications. The DenseNet and ResNet101 models both failed on the same compliant subject wearing a bright red patterned mask, bottom right in figure 4.12. It is observed that all errors in the visible band are from the deepest models. This could partially be due to overfitting on the relatively small database. These models would likely benefit from training on a larger dataset. Other misclassifications could be due to the large variability of the masks, including color, texture, and patterns that may confuse the classifier.

Thermal Results

The results for the thermal dataset in terms of classification accuracy were very high, with only one misclassification through all five folds when Alexnet is used,

Figure 4.12: Samples of compliant(top left) and non-compliant (bottom left) faces misclassified using NASNetMobile. The only thermal face misclassified by AlexNet (top right), and the face misclassified by ResNet101 and DenseNet (bottom right).

as shown in the top right corner of figure 4.12. With only one sample classified incorrectly, it is difficult to draw a conclusion about where these models struggle with the face mask classification task in the thermal band. I can, however, infer that mask compliance in the thermal spectrum is more accurate than in the visible spectrum. This is likely to due to the lack of variance of the masks themselves and how pronounced the nose is in thermal images. For thermal masks, little to no texture or color information is present, with all of them being various shades of black to dark gray. The nose in most samples is colder than the rest of the face, making it a very distinguishable feature and easy to detect if it is covered by a mask or not. Additionally, different levels of mask compliance appear to be easier to classify due to the lack of mask variation. In thermal images, a mask appears as a large dark covering over the face that is highly distinct.

Figure 4.13: Samples of misclassified images from the FMLD test set. The true labels are (A) compliant , (B) non-compliant with a mask, and (C) non-compliant with no mask.

The masks in the visible images vary in color, pattern, and texture, which adds difficulty to the final classification decision.

FMLD Test Set Results

I also evaluate the visible spectrum trained models on the FMLD test set in Batagelj et al., 2021. This data is quite different from the dataset used in the previous experiments, with a high degree of diversity in face pose, various occlusions, degree of face mask coverage, illumination, and image resolution. All results can be seen in table 4.9. It is shown that performance varied considerably, with a 33.4% difference between the best and worst performing model. SqueezeNet, ResNet101, and Alexnet performed the best, with accuracies of 89.3%, 87.3% and 85.9% respectively. Two of the top three were the shallowest models in terms of depth, with the exception being ResNet101. DenseNet also performed well.

These scores were not as good as the results in Batagelj et al., 2021, where all models achieved over 97% accuracy and only a 1.12% difference between the worst, AlexNet, and best, ResNet152, models. This was expected, as I did not train on this data and was interested in getting a better understanding of where the models fail on more challenging data. The compliant cases are much easier to label correctly, with all models achieving over 91% and many over 99%. The non-compliant cases where the mask is worn incorrectly was the most difficult, with the best model getting 77% correct and the next best only getting half of them correct. This case does highlight a need for more data because the con-

Table 4.9: Accuracy results on FMLD test set using our trained classifiers. Compliant samples account for 38% of the test set, while the non-compliant cases account for the remaining 62%. The non-compliant incorrectly worn samples are clearly the most difficult to classify for all models, with results ranging from 20.01% to 77.2% accurate.

Model	Compliant(38%)	NC-Incorrectly Worn(2.5%)	NC-No Mask(58.5%)	Total
AlexNet	0.981	0.438	0.796	0.859
SqueezeNet v1.1	0.921	0.500	0.891	0.893
ResNet18	0.992	0.407	0.482	0.678
RenNet50	0.969	0.503	0.726	0.814
DarkNet53	0.991	0.244	0.287	0.559
EfficientNet-Bo	0.993	0.201	0.315	0.575
ResNet101	0.912	0.772	0.851	0.873
DenseNet-201	0.993	0.469	0.599	0.748
NASNetMobile	0.992	0.207	0.364	0.604

compliant class accounts for 2.5% of the training and test sets. The other non-compliant cases, those with no mask present, were quite challenging. Results varied between just over 31% to almost 90%. Samples of misclassified images from all three levels of masking that I investigated are visualized in figure 4.13.

Limitations

Although the results from this evaluation are quite good on the collected thermal and visible datasets, there are a few limitations that must be mentioned. Most importantly, the dataset is quite small, using only 1,000 samples per class in each spectrum. A larger dataset with additional subjects would introduce more variance and likely more errors. This would allow for a more complete investigation into the factors that lead to a decrease in classification performance. Next, the samples are relatively easy to classify because they were collected indoors with consistent lighting at the same distance. I also did not include any profile faces, which would increase the degree of difficulty. Lastly, I ignored the face detection step in the pipeline and only assessed the classification taks using faces that have already been detected and cropped. All the mentioned limitations should not be ignored in future work.

4.1.4 MultiSpectral Face Dataset Collection

Dataset Demographics

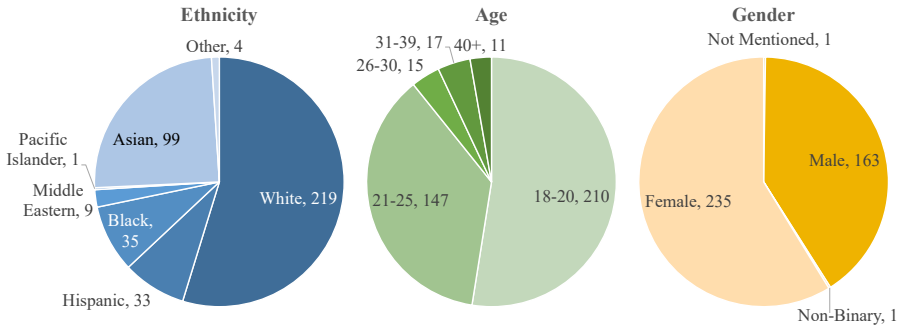


Figure 4.14: Overview of the MILAB-VTF(B) ethnicity, age, and gender demographics.

I include an overview of the demographics from the dataset after the collection activities were completed in figure 4.14. Samples of the outdoor faces and distances are shown in figure 4.15.

4.2 Keypoint Detection

4.2.1 Facial Landmark Detection

Datasets

I perform all experiments on the ARL-VTF and MILAB-VTF(B) datasets.

DEVCOM ARL-VTF

The DEVCOM ARL-VTF dataset D. Poster et al., 2021 is a large paired visible and thermal face dataset. It includes annotated landmarks (left eye center, right

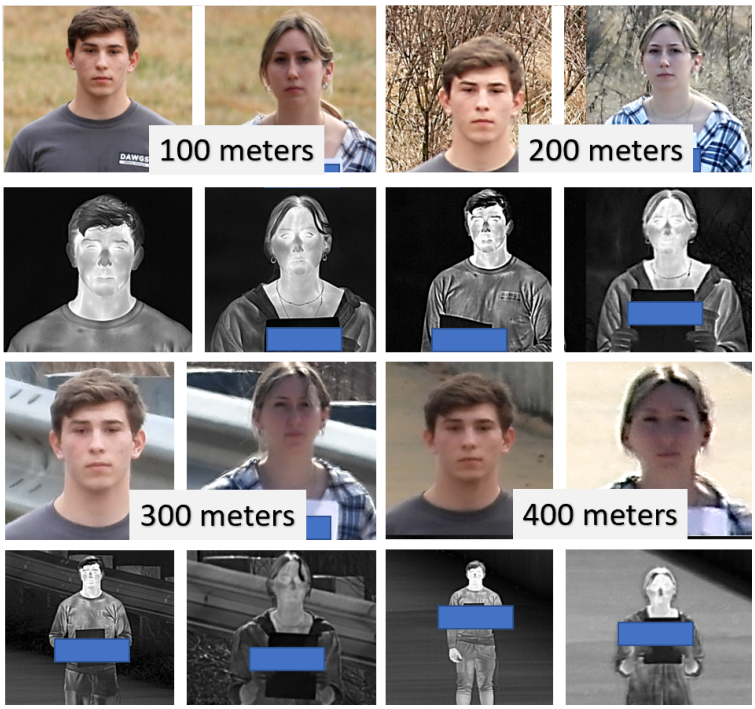


Figure 4.15: Samples of images collected for the MILAB-VTF(B) dataset outdoors in the thermal (MWIR) and visible spectrums.

eye center, nose base, left mouth corner, center of mouth, and right mouth corner) and face bounding boxes. The dataset contains 395 subjects captured at a distance of ∼2.1 meters using one LWIR and three visible cameras. Four sequences of data are captured for each subject; a baseline sequence of frontal images with a neutral expression; an expression sequence of frontal images where the subject counts out loud; a pose sequence where the subjects slowly turn their heads from left to right; and a glasses sequence where subjects who naturally wear glasses put them on and repeat the baseline sequence.

MILAB-VTF(B)

The MILAB-VTF(B) dataset was first presented in Peri et al., 2021a. It is a very challenging multispectral, multi-distance face dataset. Unsynchronized, paired data of 400 subjects were captured indoors at a distance of 1.5 meters, and outdoors at distances of 100-, 200-, 300-, and 400-meters. A Canon Mark IV and Nikon P1000 were used to capture the indoor and outdoor data, respectively. The MWIR data was captured with a FLIR A8581 indoors and a FLIR RS8513 outdoors. Each sequence recorded the subject turning their head left-and-right

and up-and-down for approximately 30 seconds. Annotations for the dataset include face bounding boxes and 21 facial landmarks using the method in Peri et al., 2021a. An additional subset of the data includes five images of each subject at every distance and spectrum, manually labeled for face bounding boxes and seven facial landmarks. The landmarks include the left and right eye corners, tip of the nose, and left and right mouth corners.

Evaluation Metrics

I use the NRMSE, the standard evaluation metric Zafeiriou et al., 2017 for reporting landmark detection performance. The formula for NRMSE is

$$E(\hat{L}, L) = \frac{\|\hat{L} - L\|_2}{N_{scale}} \tag{4.2}$$

where L and $\hat{L}$ are the ground truth and predicted shapes respectively, $\|.\|_2$ is the ℓ_2 norm, and N_{scale} is the normalization factor. I also choose the face diagonal for the normalization factor as in Zafeiriou et al., 2017 and D. Poster et al., 2021 due to its ability to handle changes in face pose better than the inter-ocular distance.

In addition to the mean error, I also report standard deviation (STD), median, Median Absolute Deviation (MAD), and maximum error. Area under the curve (AUC) and failure rates are set to a threshold of 0.08. Failure rate is the percentage of images where the average predicted landmark errors exceed the chosen threshold. AUC and failure rates are visualized in Cumulative Error Distribution (CED) curves as well as being reported in tables 4.10, 4.12, and 4.13.

Synthesis Details

StarGAN v2 and CUT: In this processing module I fine-tuned StarGAN v2 and CUT on the ARL-VTF dataset for the two classes, visible and thermal faces. Over 8,000 images, split evenly between all subjects and scenarios, were used during training. For StarGAN v2, I found that the default hyperparameters from Choi et al., 2020 resulted in the best synthesis results. The model was trained up to 150,000 steps, and I observed that 100,000 steps yielded the best model in terms of landmark detection accuracy and is subsequently used for all reported results. CUT was trained to 115 epochs, using the hyperparameters from Park et al., 2020. StarGAN v2, CUT, and HRNet both require input dimensions of 256×256, therefore, all face data was preprocessed by cropping and scaling the faces in order to match the face images used in Choi et al., 2020.

Then, the ground truth facial landmark locations were transformed to match the cropped and scaled faces. Finally, face images were synthesized with both models. After synthesis, facial landmarks could be predicted and evaluated without requiring any further processing.

Samples of ground truth thermal and visible faces, and faces synthesized by StarGAN v2 and CUT are visualized in figure 4.16. Both methods use the visible ground truth faces as a reference for synthesis. I observed that, in general, StarGAN v2 generates more realistic looking faces, but both models do a good job of preserving the source characteristics of the face, especially the face pose, eyes, and nose locations.

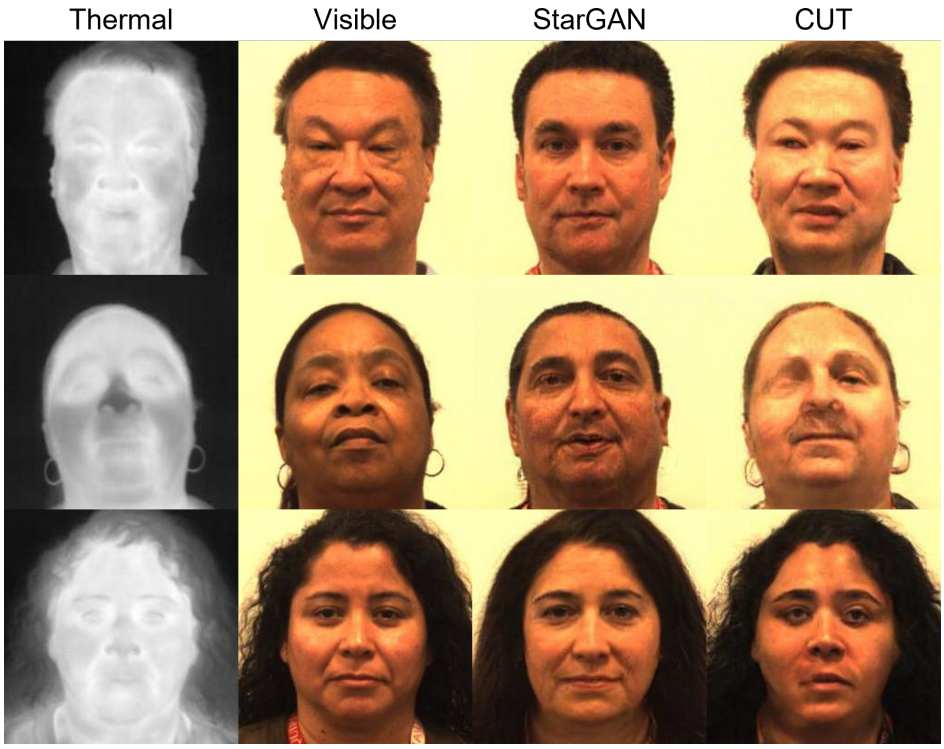


Figure 4.16: Comparison of the StarGAN v2 and CUT methods for thermal-to-visible face synthesis on the ARL-VTF dataset.

Baseline Synthesis and Landmark Detection

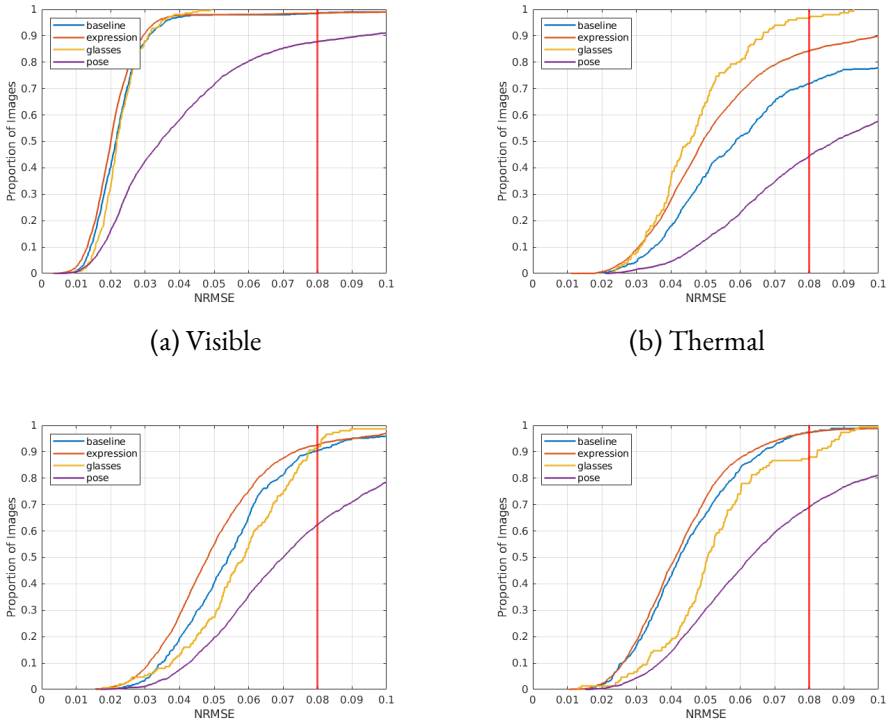
A baseline HRNet model, pre-trained on visible data from the WFLW dataset, was used to investigate the performance gap between landmark predictions on visible and thermal data using the ARL-VTF dataset. The detection errors from the thermal data were much higher, as expected, especially in the pose scenario. The same HRNet model was used for landmark prediction on latent-guided and

Table 4.10: Summary of results using an HRNet model pre-trained on the WFLW dataset. Baseline results from the thermal and visible bands of the ARL-VTF database show the performance gap between the two spectrums. By synthesizing the thermal data using latent and reference-guided synthesis, results improve closer to visible band performance.

Scenario	Mean	Std	Median	MAD	Max Error	$AUC_{0.08}$	Failure $Rate_{0.08}$
Visible							
Baseline	0.0241	0.0197	0.0215	0.0042	0.2390	0.7137	0.0157
Expression	0.0223	0.0160	0.0200	0.0040	0.2194	0.7311	0.0144
Pose	0.0480	0.0458	0.0344	0.0134	0.5874	0.4990	0.1228
Glasses	0.0229	0.0063	0.0218	0.0036	0.0491	0.7141	0
Thermal							
Baseline	0.0779	0.0551	0.0587	0.0174	0.3639	0.2655	0.2820
Expression	0.0597	0.0381	0.0491	0.0123	0.4421	0.3484	0.1572
Pose	0.1260	0.1057	0.0873	0.0336	1.2443	0.1228	0.5548
Glasses	0.0472	0.0142	0.0456	0.0077	0.0928	0.4136	0.0333
StarGAN v2 Latent-Guided Synthesis							
Baseline	0.0569	0.0227	0.0540	0.0103	0.2012	0.3208	0.0960
Expression	0.0520	0.0225	0.0481	0.0102	0.3224	0.3769	0.0740
Pose	0.0815	0.0461	0.0700	0.0192	0.4885	0.1805	0.3764
Glasses	0.0582	0.0162	0.0587	0.0111	0.1068	0.2791	0.0867
StarGAN v2 Reference-Guided Synthesis							
Baseline	0.0455	0.0183	0.0424	0.0102	0.1555	0.4417	0.0260
Expression	0.0437	0.0173	0.0413	0.0094	0.1716	0.4630	0.0268
Pose	0.0740	0.0427	0.0629	0.0188	0.4009	0.2351	0.3096
Glasses	0.0531	0.0172	0.0505	0.0085	0.1013	0.3478	0.1200
CUT Synthesis							
Baseline	0.0391	0.0158	0.0364	0.0081	0.1268	0.5185	0.0260
Expression	0.0381	0.0163	0.0343	0.0078	0.1626	0.5299	0.0284
Pose	0.0691	0.0465	0.0566	0.0178	0.4931	0.2854	0.2560

reference-guided synthesized images from StarGAN v2. For reference-guided synthesis, I used the same subject for the thermal source and visible reference images. As seen in table 4.10, there is a clear performance advantage to using a reference image to perform synthesis as opposed to latent codes. Samples of these predictions can be seen in figure 4.18. The CED plots in figure 4.17 show that the pose scenario remains the most challenging across all spectrums and synthesis methods. The baseline and expression scenarios had the lowest error on the synthesized data, with the expression data being slightly lower than the baseline. The difference in error is likely due to the more distinguishable mouth corners in the expression data when compared to the baseline, as the only difference between the two is the subject opening their mouth. I also compared CUT to StarGAN v2 reference-guided synthesis in table 4.10. I observed im-

provement in landmark detection performance using the CUT model in the baseline, expression, and pose scenarios. I discuss the reasons for this improvement and analyze errors in more detail in the next sections.



(a) Visible

(b) Thermal

(c) Latent-Guided

(d) Reference-Guided

Figure 4.17: CED curves for visible (a), thermal (b), latent-guided (c), and reference-guided (d) synthesis on the ARL-VTF dataset using StarGAN v2. Failure threshold is shown at 0.08 (red line).

Error Analysis

I attribute the majority of landmark detection errors to two major factors. The first is the synthesis process and how well the source characteristic locations are preserved. The second is incorrect ground truth labels in the ARL-VTF dataset. It is important to note that the ARL-VTF dataset was not manually annotated.

Table 4.11 reports the mean error of each landmark from the baseline and expression scenarios using visible faces and the StarGAN v2 and CUT synthesized faces. These results illustrate two important observations.

- First, it helps confirm my previous assumption that the mouth corners are easier to detect in the expression scenarios because the greatest decrease in landmark error between baseline and expression occurs at both mouth corners.

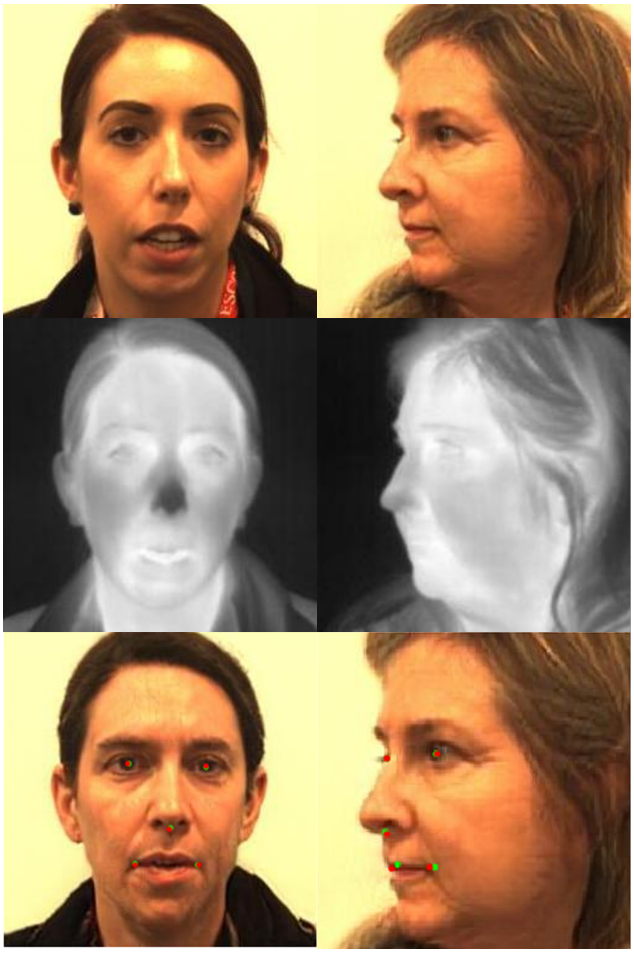


Figure 4.18: Examples of accurate thermal-to-visible synthesis and landmark detections using reference-guided synthesis on ARL-VTF dataset. Visible reference (top), source thermal (middle), and (bottom) StarGAN v2 synthesized faces with ground truth (green, from the method used in D. Poster et al., 2021) and predicted (red) landmarks.

- Second, it reveals that the synthesis process projects most of its error during image generation onto the mouth corners more than the other landmarks.

When I take the mean error for the left and right eyes and compare them to the mean error of the left and right mouth corners of the visible data, the mouth corners have a 25.0 % and 16.7% higher error for the baseline and expression scenarios respectively. For the StarGAN v2 synthesis landmarks, errors are 41.8% and 23.9% higher for the baseline and expression scenarios. This same pattern

holds true for the CUT data, confirming that the mouth corners are more difficult to accurately synthesize, especially for the baseline scenario. While some of this error may be unavoidable during the synthesis process, I believe that much of the error can be explained by the challenging nature of the thermal images in the dataset.

In figure 4.19 I show that it is often quite difficult to visually detect exactly where the mouth corners begin when compared to the visible data. I find the case on the left side of figure 4.19 to be more common than the case on the right side, where the mouth is more clearly defined. As table 4.11 confirms, instances where the mouth is open are often easier to synthesize the mouth corners more precisely than when the mouth is closed.

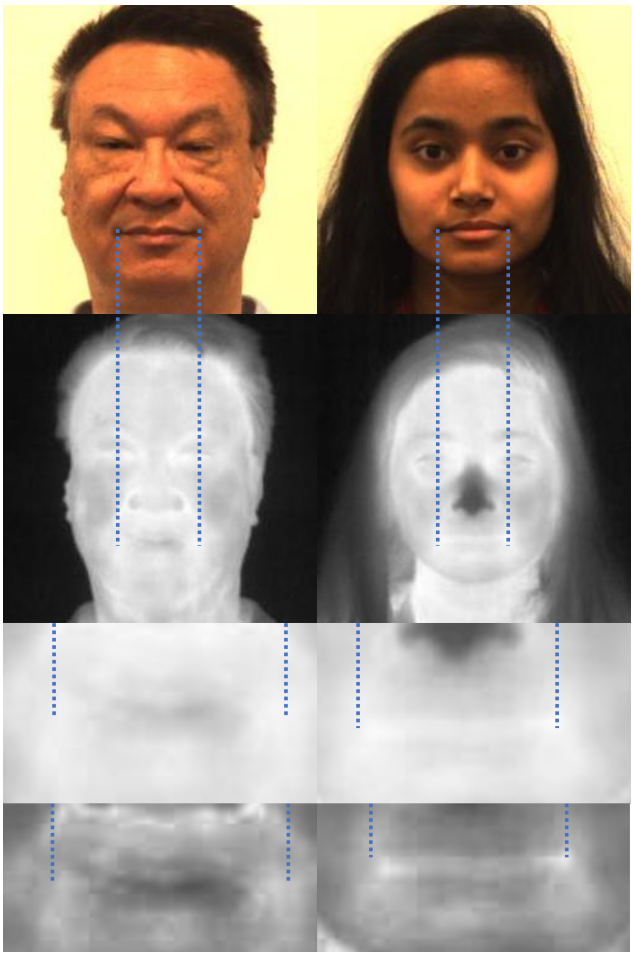


Figure 4.19: Comparison of mouth corner visibility from two subjects in the ARL-VTF dataset. The mouth corners are easily seen in the visible images (top). Row 3 is the thermal face, cropped and zoomed in on the mouth. The bottom row is a contrast-enhanced look at the same mouth, further illustrating how some mouth corners can be more easily seen than others.

Table 4.11: ARL-VTF mean error by landmark for the baseline and expression scenarios comparing visible faces to StarGAN v2 reference-guided and CUT synthesized faces.

Scenario	Left Eye	Right Eye	Left MC	Right MC	Nose Base
Visible Baseline	0.0211	0.0213	0.0274	0.0255	0.0249
Visible Expression	0.0205	0.0203	0.0240	0.0235	0.0230
StarGAN v2 Baseline	0.0391	0.0364	0.0531	0.0541	0.0444
StarGAN v2 Expression	0.0420	0.0365	0.0489	0.0485	0.0422
CUT Baseline	0.0361	0.0318	0.0507	0.0449	0.0319
CUT Expression	0.0387	0.0351	0.0439	0.0400	0.0327

I further examine the landmark detection errors with figure 4.20. After checking images where the error was greater than the 0.08 failure threshold, I observed a significant number of images, especially those with errors closest to the maximum error from the dataset, were simply due to inaccurate ground truth labels. I show in figure 4.20 a few samples of images where the synthesis and landmark predictions appear to be very accurate, but the ground truth labels are not. Even when I factor in the previously discussed errors, where small inaccuracies during the synthesis process created mouth corners in slightly different positions than the locations in the ground truth thermal face, the ground truth landmarks are still very inaccurate. After visually checking all images for errors created by reference-guided synthesis in the baseline, pose, and expression scenarios, I estimate about 18% of all images with errors over the failure rate threshold have ground truth coordinates that can easily be observed as wrong for multiple landmarks per face. In these cases, the predicted landmarks are shown to be highly accurate nearly 100% of the time.

The error analysis leads to a few interesting observations. First, the synthesis of both methods nearly always produces a realistic looking face with the source characteristics preserved to an acceptable degree for facial landmark detection. Second, the HRNet landmark detector seems to work very well for nearly all synthesized faces, even in cases where the synthesis process does not produce a realistic looking face. As I learned through our visual inspection of errors that were over the failure threshold, it is rare for the landmarks to be predicted in the wrong location. The majority of images with the largest errors come from incorrect ground truth labels, while the other errors are mostly due to the slight offset between ground truth source characteristic positions and their synthesized positions, most commonly the mouth corners. Aside from accurate ground truth annotations, I believe the most important part of the landmark detection process is during the training phase of the synthesis model, where the cyclic loss is key to preserving the correct locations of the source characteristics

Figure 4.20: Examples of detections from StarGAN v2 synthesized ARL-VTF faces where error is over the 0.8 failure threshold. In many cases using our method, the predictions (red) appear accurate, but calculated as over the threshold because of incorrect ground truth annotations (green).

on the synthesized faces. I hypothesize that CUT was able to outperform StarGAN v2 in terms of landmark detection, even with less visually appealing faces, because it was better able to preserve the locations of the eyes, nose, and mouth corners. This may have been due to CUT's patch-based approach to translation or better convergence during training than that of StarGAN v2.

Comparison with State-of-the-Art

I implemented transfer learning to train two HRNet models for comparison against current state-of-the-art results reported on the ARL-VTF dataset. When

training HRNet, I found that the learning rate made the largest difference in performance, and an ideal learning rate was found for each dataset and spectrum. The first model attempts to transfer knowledge already learned in the visible domain to the thermal domain by training with only thermal data. The second model shows the ability of HRNet to learn the domain invariant features shared between the visible and thermal models by training with the entire dataset, using visible and thermal images and all four sequences of data. Results from these models are only reported for the thermal portion of the test set. Results are shown in table 4.12, comparing my two methods with the results reported in D. Poster et al., 2021 and Peri et al., 2021b.

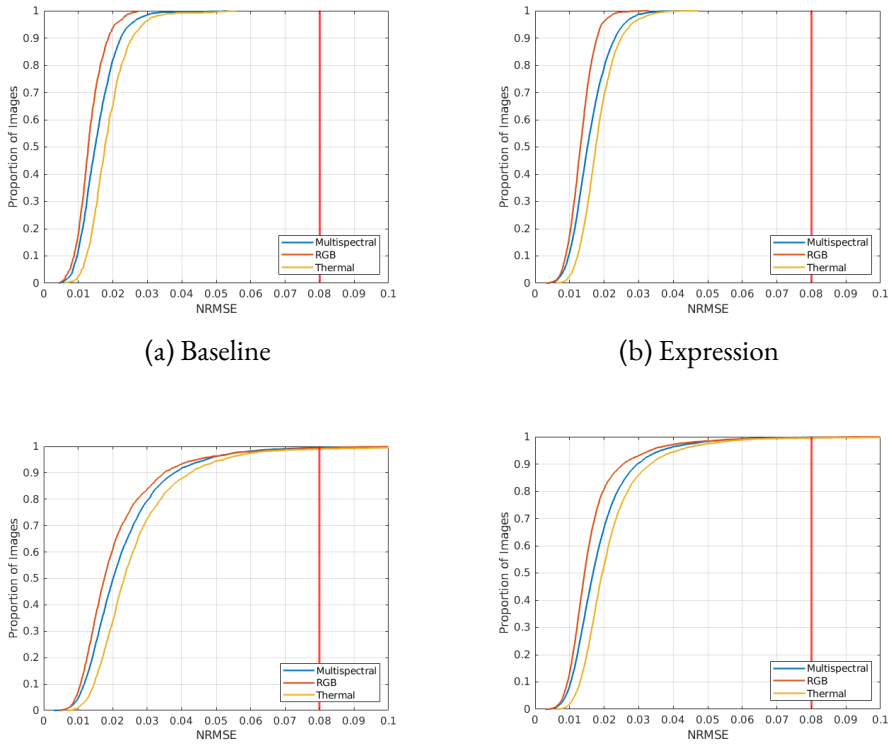
Table 4.12: ARL-VTF landmark evaluation results. Our models trained on thermal and multispectral (thermal and visible) data both perform better than previous methods.

Scenario	Method	Mean	STD	Median	MAD	Max Error	$AUC_{0.08}$	Failure $Rate_{0.08}$
Baseline	DAN D. Poster et al., 2021	0.0326	0.0155	0.0283	0.0119	0.0857	0.5798	0.0080
	Novel Faster-RCNNPeri et al., 2021b	0.0192	0.0069	0.0179	0.0037	0.0631	0.7606	**0.0**
	Ours (Thermal)	**0.0186**	0.0059	**0.0177**	0.0032	0.0554	**0.7672**	**0.0**
	Ours (Multispectral)	0.0191	**0.0056**	0.0186	**0.0031**	**0.0535**	0.7609	**0.0**
Expression	DAN D. Poster et al., 2021	0.0324	0.0157	0.0276	0.0122	0.1109	0.5946	0.0076
	Novel Faster-RCNNPeri et al., 2021b	0.0212	0.0095	0.0190	0.0047	0.1106	0.7345	0.0010
	Ours (Thermal)	**0.0183**	0.0052	**0.0176**	**0.0031**	0.0472	**0.7715**	**0.0**
	Ours (Multispectral)	0.0189	**0.0050**	0.0183	0.0033	**0.0440**	0.7640	**0.0**
Pose	DAN D. Poster et al., 2021	0.1012	0.0562	0.0949	0.0472	0.4431	0.1692	0.5868
	Novel Faster-RCNNPeri et al., 2021b	0.0316	0.0186	0.0265	0.0086	0.2145	0.6116	0.0290
	Ours (Thermal)	0.0267	0.0140	**0.0233**	0.0061	0.1700	0.6691	0.0100
	Ours (Multispectral)	**0.0264**	**0.0124**	0.0235	**0.0058**	**0.1272**	**0.6704**	**0.0056**

Both of my models achieved better performance across all reported metrics when compared to the other methods discussed in the recent literature. Results for the baseline and expression scenarios were very similar, and slightly better performance overall can be observed in the model trained on multispectral data for the pose scenario. Figure 4.21 shows CED curves comparing both of my models against a baseline model trained only on the ARL-VTF visible data using the baseline, expression, and pose scenarios. An average of all scenarios is also shown in figure 4.21d.

Evaluation on MILAB-VTF(B)

Baseline landmark detection results are presented for the manually annotated portion of the MILAB-VTF(B) dataset. I used transfer learning with a pre-trained HRNet model that predicted 19 points from the WFLW dataset onto the five landmarks of the ARL-VTF training set. Then, I train this model again on MILAB-VTF(B)'s training set of seven landmarks before evaluating on the test set. Training and testing was performed using the thermal and visible data



(a) Baseline

(b) Expression

(c) Pose

(d) All

Figure 4.21: CED curves for landmark detection models trained on visible, thermal, and multispectral data from the ARL-VTF dataset. Failure threshold is shown at 0.08 (red line). Sequences include baseline (a), expression (b), pose (c), and the mean of all sequences (d).

from all five distances. Default hyperparameters used in J. Wang et al., 2020 were used throughout the entire process. Results in table 4.13 indicate that MILAB-VTF(B) is much more challenging than the ARL-VTF dataset for landmark detection. While the indoor and 100-meter error is only slightly higher than ARL-VTF, the 200-, 300-, and 400-meter distances quickly degrade in performance, especially in terms of AUC and failure rate. I observed that the mean results for the 100-meter data are actually better than the indoor scenario. This is likely due to the high resolution and zoom ratio of the thermal and visible cameras that were used during the creation of the outdoor portion of the dataset, making the face images look similar in quality to that of the data captured indoors.

Table 4.13: MILAB-VTF(B) baseline multispectral landmark evaluation.

Distance	Mean	STD	Median	MAD	Max Error	$AUC_{0.08}$	Failure $Rate_{0.08}$
Indoor	0.0265	0.0247	0.0222	0.0055	0.4666	0.6883	0.0125
100 Meter	0.0260	0.0167	0.0231	0.0061	0.2478	0.6840	0.0100
200 Meter	0.0322	0.0237	0.0283	0.0077	0.4235	0.6155	0.0200
300 Meter	0.0391	0.0321	0.0339	0.0104	0.5561	0.5393	0.0451
400 Meter	0.0472	0.0307	0.0401	0.0149	0.3512	0.4497	0.0990
All	0.0342	0.0274	0.0280	0.0088	0.5561	0.5955	0.0373

4.3 Feature Extraction and Matching

4.3.1 Race and Gender Classification for Cross-spectral Face Recognition

Dataset and Metrics

I use the MILAB-VTF(B) dataset from section 1.2.4 for all experiments in this section. I do not use any outdoor data, and instead focus only on the baseline indoor images. All face images have been preprocessed for detected faces and geometrically aligned using the ground truth eye locations. The training set was composed of 2,770 images and the test set was composed of 138 images. Subjects were separated into train and test sets using the MILAB-VTF(B) protocol. All face recognition experiments contained one image in the gallery set and one image in the probe set. As seen in figure 4.22, the train and test sets are balanced in terms of gender, but highly imbalanced in terms of race, with White being nearly 60% of the dataset, Black almost 11%, and Asian 28%. Because of the data imbalance, a multitask learning model for gender and race classification was selected due to the models ability to efficiently use data and reduce overfitting Crawshaw, 2020.

I used standard classification and face recognition metrics for all experiments. I report precision, recall, and F1 scores for gender and race classification. I evaluated results for the verification and identification tasks using standard metrics Phillips et al., 2011. I plot the ROC curve for verification metrics and report the EER and AUC. For identification metrics I plot the Cumulative Match Characteristics curve and report rank scores.

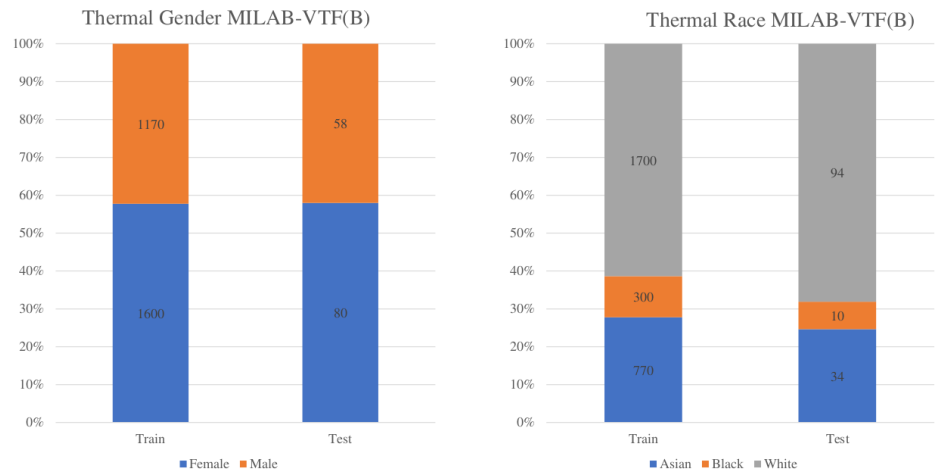


Figure 4.22: Summary of the training and test sets from MILAB-VTF(B). Classes are balanced for gender (left) and imbalanced for race (right).

Race and Gender Classification

It is clear from the results that classification of race and gender in MWIR is very challenging. Race classification is especially difficult. However, in table 4.14, results show that the multitask network learned to classify all visible spectrum classes successfully, with the lowest F1 score of 90.9% for the Black race and a high of 98.9% for the White class. These results reflect the imbalance of the dataset, but show that overfitting was avoided.

Table 4.14: Classification results from a multitask network trained with visible images.

	Precision	Recall	F1	Count
Asian	0.941	0.941	0.941	34
Black	0.833	1.000	0.909	10
White	1.000	0.978	0.989	94
Female	0.939	0.975	0.957	80
Male	0.963	0.913	0.938	58

The MWIR results in tables 4.15 and 4.16 show the two best performing multitask classification models. The table 4.15 model generalized better to the gender classes and table 4.16 model generalized better to the race classes. The

results on the Asian class are clearly the most difficult, with a high F1 score of 45.2% from the model reported in table 4.16 and a low F1 score of 30.7% from the model reported in table 4.15.

Table 4.15: Classification results from the multitask network trained with MWIR images that performed better on the gender classes.

	Precision	Recall	F1
Asian	0.444	0.235	0.307
Black	1.000	0.700	0.823
White	0.761	0.914	0.830
Female	0.975	1.000	0.987
Male	1.000	0.965	0.982

Table 4.16: Classification results from the multitask network trained with MWIR images that performed better on the race classes.

	Precision	Recall	F1
Asian	0.500	0.412	0.452
Black	1.000	0.900	0.947
White	0.792	0.851	0.821
Female	0.897	0.975	0.934
Male	0.961	0.845	0.899

The results using the EfficientNet model also confirm that race classification is very difficult in the MWIR band, with similar results to the multitask networks shown in table 4.17. This table shows the results of training two EfficientNet models, one for the gender task and one for the race task. When training a single model for all classes, results were poor and have not been included in any of the reported results.

The predictions for race and gender are used for synthesizing face images to perform face recognition in the next section. In the face recognition scenario where the multitask network is used for gender and race prediction, the gender classification branch from the model in table 4.15 is used for predicting gender and the race classification branch from the model in table 4.16 is used for the

Table 4.17: Classification results from the EfficientNet models trained with MWIR images. One model shows results from the race classes, the other shows results from the gender classes.

	Precision	Recall	F1
Asian	0.500	0.382	0.433
Black	0.750	0.600	0.667
White	0.779	0.862	0.818
Female	0.962	0.950	0.956
Male	0.932	0.948	0.940

race predictions as these were the most accurate for each task. The EfficientNet face recognition scenario also uses the two models reported in table 4.17.

Cross-spectral Face Recognition

Results for all face recognition scenarios using ArcFace and VGGFace are shown in tables 4.18 and 4.19, respectively. The visible to visible matching scenarios for both matchers are almost perfect, as expected. The thermal-to-visible scenario illustrates the gap in performance when using thermal images instead of visible images to match against a visible image gallery. Figures 4.23 and 4.24 show the CMC and ROC curves for the thermal-to-visible matching, respectively. The third row, where the same subject is used as the reference image for the synthesis process, shows the current best performance that can be achieved with my methods. This is the ideal scenario where the correct identity, gender, and race are used for synthesizing a visible face from a thermal face. The ArcFace matcher performed well, with a 47.8% rank-1 accuracy and 95.6% rank-10 accuracy for the identification task, and almost 90% AUC for the verification task. For an unknown reason the VGG-Face matcher reported much lower identification scores compared to ArcFace in this scenario.

The other scenarios, where race and gender are selected to be incorrect for one or both classes, performance drops significantly. All scenarios where any combination of race and gender are incorrect for the reference image before synthesis show a large drop in performance when compared to the ideal scenario where the same identity is used for synthesis and matching. From the VGG-Face experiments, it's clear that choosing the wrong race is the most con-

sequential factor face recognition performance. The effect was less noticeable using ArcFace.

Table 4.18: All ArcFace recognition results. V-V is the visible to visible matching scenario. T-V is the thermal-to-visible matching scenario.

Scenario	Rank-1	Rank-10	AUC	EER
V-V	100	100	99.73	2.92
T-V	0.00	20.29	52.32	46.57
Same Ref.	47.82	95.65	88.71	19.29
Random R-G	1.45	24.64	59.93	40.89
Wrong Gender	0.00	24.64	58.14	44.19
Wrong Race	0.00	23.19	56.02	46.14
EffNet	4.35	30.43	58.83	44.44
Multitask	7.25	33.33	60.45	42.06

Table 4.19: All VGG-Face recognition results. V-V is the visible to visible matching scenario. T-V is the thermal-to-visible matching scenario.

Scenario	Rank-1	Rank-10	AUC	EER
V-V	100	100	100	0
T-V	4.34	23.19	55.77	45.39
Same Ref.	21.73	59.42	85.11	23.78
Random R-G	1.45	24.64	62.75	41.13
Wrong Gender	7.25	42.03	74.07	37.14
Wrong Race	5.79	21.74	59.15	41.48
EffNet	8.69	36.23	69.24	34.77
Multitask	7.25	39.13	68.59	37.54

When I use either of my trained race and gender classifiers to choose the reference image for synthesis, there is a significant boost in identification and verification performance. In many cases, the multitask network was the more accurate classifier and led to better recognition performance overall. Figures 4.25 and 4.26 show the identification and verification performance, respectively. Although there remains a gap between the proposed method and the ideal scenario where the reference image is the same identity as the face being matched, knowing the correct demographic information of the MWIR face image is helpful for improving cross-spectral face recognition performance.

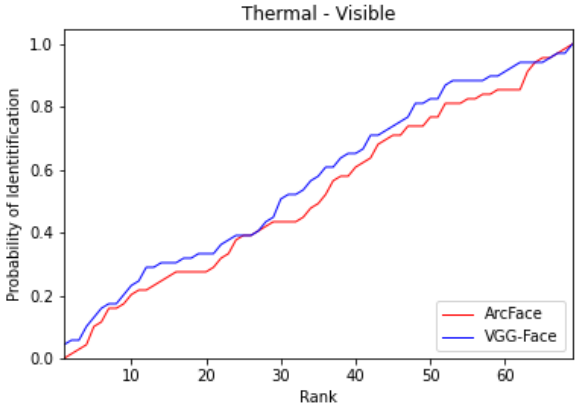


Figure 4.23: Identification results from the CMC curve matching thermal faces to visible faces.

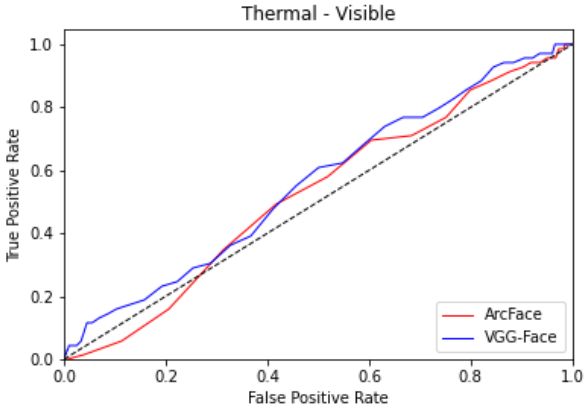


Figure 4.24: Verification results from the ROC curve matching thermal faces to visible faces.

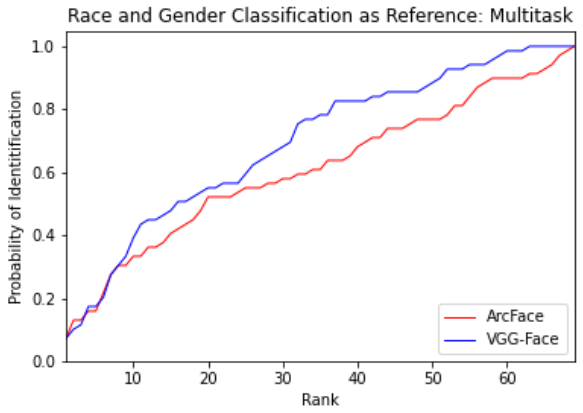


Figure 4.25: Identification results from the CMC curve using the multitask classifier to choose the reference image when matching synthesized faces to visible faces.

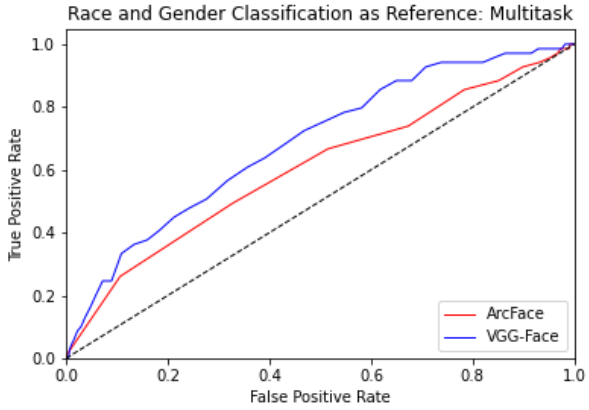


Figure 4.26: Verification results from the ROC curve using the multitask classifier to choose the reference image when matching synthesized faces to visible faces.

4.4 Other Applications

4.4.1 Firearm Detection

Data

Table 4.20 contains an overview of the dataset that was collected in this work, showing the number of collected frames and the number of frames used for all experiments after balancing the dataset and filtering our very poor quality data. In figure 4.27, there is also a summary of the video resolutions, which vary from 360p to 1080p.

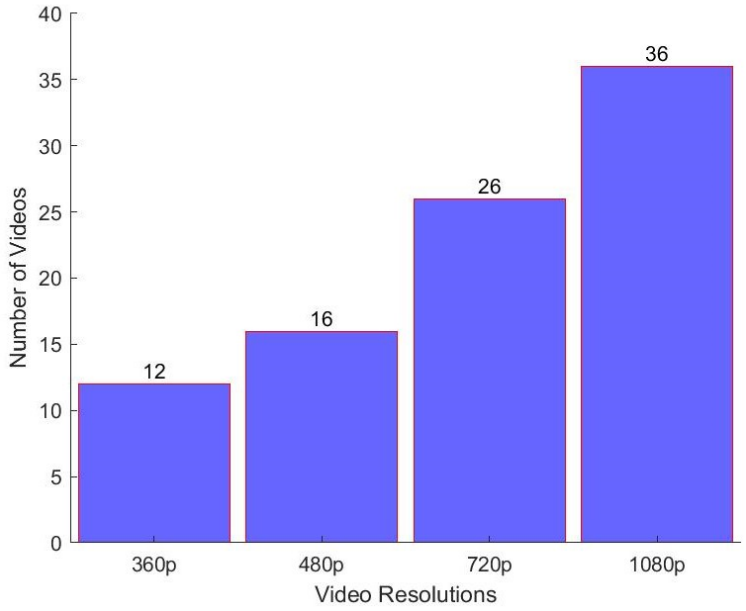


Figure 4.27: Count of the different video resolutions in the dataset.

Although over 11,000 images were collected and labeled, not all of them could be used for the experiments for two reasons. First, a large portion of the labeled data was determined to be very challenging and proved detrimental to the training process due to harsh lighting or shadows, blur from videos captured with low resolution or frame rate, very small class instances (mostly identified in the handgun class), and occlusion. Second, the dataset was imbalanced in terms of the number of usable images per class from each video. Some videos had less than 15 images where a firearm is present, while others had hundreds. In order to balance the final dataset that was uded in all experiments, I take no more than 55 images from videos containing handguns and no more than 150

images from videos containing long guns. This yielded a final dataset with 3,210 handgun images and 2,993 long gun images, totaling 6,203.

Table 4.20: An overview of our dataset showing the number of collected frames and the number of frames used for all experiments after balancing the dataset and filtering out very poor quality data.

# Videos	Handgun	Long Gun	Total
90	5,773	5,879	11,652
# of Samples Used for Experiments			
90	3,210	2,993	6,203

Detection Model and Metrics

Three sets of experiments are performed in this work. The first two experiments evaluated the data augmentation techniques that resulted in better object detection performance, while the last experiment optimizes the hyperparameters of the final model using those techniques. For all experiments, and after experimenting with other architectures, I chose a Faster R-CNN architecture using the ResNet-50 base network, previously trained on the COCO dataset. Using transfer learning, I trained this model to detect the two firearm classes. I report precision and recall at a 0.5 intersection over union (IoU) threshold, and the mean average precision $mAP^{0.5:.0.95}$, where mAP is averaged over 10 IoU thresholds. Additionally, I report F1 scores for each of the two classes and plot the true detection rate (recall) vs the false alarm rate, at different IoU thresholds. All reported results use a confidence score threshold of 0.5.

Selecting Augmentation Techniques

Due to the large number of experiments needed to assess all 18 data augmentations, I split the dataset into three non-overlapping folds with 90% of the data used for training and 10% for testing. I used 3-fold cross-validation in the first two experiments to avoid overfitting and also to train all models in a reasonable amount of time. The data for each fold in every experiment was selected randomly and avoided overlapping samples in the test folds. In the first experiment, I trained a baseline model with the Faster R-CNN architecture that was previously trained on the COCO dataset. Using that model's default hyperparameters, I retrained it on my data without any data augmentations. The resulting model was used as the baseline. The baseline model was trained

using stochastic gradient descent with a learning rate of 0.0001, momentum value of 0.9, and 8 epochs. Then, I trained the model using cross-validation for each of the eighteen data augmentations in order to identify which methods improved or impaired firearm detection performance. Once this evaluation was completed, I identified the best 5 performing augmentations. I used these for the next experiments to retrain the Faster R-CNN on all combinations of these 5 augmentations, totaling 26 different trained models. I use only 5 augmentations for the augmentation experiments because of time constraints. If 6 augmentations had been chosen, this decision would have resulted in 57 combinations. Once completed, I choose the augmentation combination that yielded the highest firearm detection performance and used it in final experiment.

Hyperparameter Optimization and Proposed Model

For this experiment I performed hyperparameter optimization and empirically evaluated the impact of different optimization algorithms on firearm detection performance. First, I used the most accurate object detection model that was determined from the previous experiment and found the optimal hyperparameters. This process resulted in the final proposed object detection model. Hyperparameter tuning was achieved by varying the epochs and learning rate of the model and evaluating three different optimization algorithms; (1) stochastic gradient descent with momentum Qian, 1999, (2) RMSprop Tieleman and Hinton, 2012, and (3) ADAM Kingma and Ba, 2014. Once the optimization training was completed using 8 epochs, I selected the best optimization algorithm and tested another range of learning rates. I repeated the previous training strategy except I used a smaller range of learning rates with 10-fold cross-validation to complete the optimization process.

Augmentation Experiments

From the 18 augmentation techniques I assessed, nine were found to improve performance and nine decreased or had little effect on performance in terms of mean average precision, precision, recall, and F1 scores. Of those that improved performance, I chose random pixel value scale, image scale, brightness, hue, and saturation changes as techniques to evaluate further. I trained models for all combinations of these five augmentation techniques to determine which of them should be used for optimizing and training the final model. After comparing the results, I chose a combination of four out of the five augmentations. The random pixel scale technique was not used in the final optimization experiments, although there were several other combinations that had comparable

scores to the one I chose. These scores were compared against the baseline model using no augmentations in table 4.21. When compared to the baseline, there is an almost 13% increase in $mAP^{0.5:.0.95}$. The detection accuracy of handguns increased significantly, especially the precision, recall, and F1 scores. The detection accuracy of long guns also improved, but marginally when compared to the handgun class. This is not surprising because, as I empirically determined, the long gun class has image samples that are larger in terms of spatial resolution and are often much better in terms of image quality, resulting in more discriminating features when compared to the image samples of the handgun class.

Table 4.21: Results of our baseline detection model that used no augmentation techniques compared with optimized models that used augmentations. The performance increase of our best model in terms of $mAP^{0.5:.0.95}$, precision, recall, and F1 score are provided for each class in the last column. All metrics use an IoU of 0.5 except for $mAP^{0.5:.0.95}$. Model 1 uses pixel scale and hue augmentations. Model 2 uses brightness, hue, and saturation.

Detection Results					
Metric	Baseline	Model 1	Model 2	Final	% Change
$mAP^{0.5:.0.95}$	0.541	0.554	0.565	0.672	+ **13.1**
Handguns					
Precision	0.820	0.863	0.890	0.939	+ **11.9**
Recall	0.836	0.813	0.793	0.964	+ **12.8**
F1	0.828	0.887	0.889	0.951	+ **12.3**
Long Guns					
Precision	0.873	0.893	0.877	0.952	+ **7.9**
Recall	0.923	0.920	0.917	0.946	+ **2.3**
F1	0.897	0.906	0.896	0.949	+ **5.2**

Final Optimization Experiments

After determining the best combination of augmentation techniques, I jointly assess the learning rate and optimization method for three different algorithms, stochastic gradient descent with momentum, ADAM, and RMSprop. I found RMSprop to yield the highest performance in terms of detection accuracy. After training again on a smaller range of learning rates that performed best, the final model was trained using RMSprop with a learning rate of 0.0008 and a

momentum value of 0.9 for a total of 16 epochs. Figure 4.28 plots the true detection and false alarm rates for the handgun and long gun classes using an IoU of 0.5 and 0.9.

Acceptance Criteria

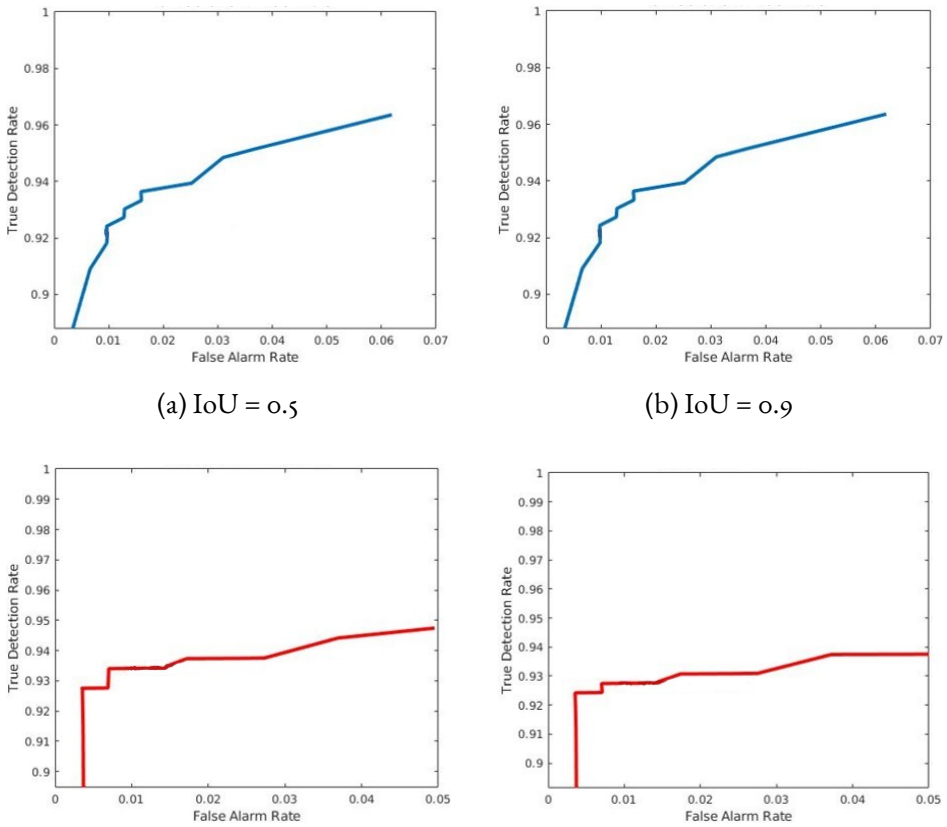
This work required the creation of a new dataset. With the creation of this dataset, an acceptance criteria of target metrics was set at the beginning of this work by the project sponsor. This criteria included:

- Expected frame rate for processing detections between 7-15 frames per second (fps).
- 80% or better True Detection Rate at 1% False Alarm (False Match) Rate for small arm detection under direct sunlight illumination or artificial spectrum illumination at distances of up to 15 meters for handguns and up to 75 meters for long guns.

The expected frame rate criteria was achieved with a reported rate of 10 fps using an Nvidia Titan Xp with 12 GB of memory and a 1.42GHz clock speed. For the detection criteria, a True Detection Rate of 92.16% at 1% False alarm rate was reported for the handgun class and a True Detection Rate of 93.41% at 1% False alarm rate was reported for the long gun class. Each of these metrics were well above the expected target metric of 80%. All detections were evaluated with the Intersection over Union (IoU) metric set at 0.5, a common threshold for a positive detection. The IoU is an evaluation metric where the area of overlap between the predicted and ground truth bounding boxes is divided by the area of union of the predicted and ground truth bounding boxes. The higher the IoU threshold is set, the more precise the predicted bounding boxes are for a positive detection. I also evaluated the weapon detector at an IoU threshold of 0.9, and observed detection rates of 92.15% and 92.7% for the hand gun and long gun classes, respectively.

Summary of Results

The plots at each of the IoU thresholds are very similar, meaning the bounding box predictions in most cases remain very accurate as the IoU threshold increases, all while maintaining a similar detection and false alarm rate. This performance can partially be attributed to the nature of the objects that are being detecting. In most cases, the weapon is a very small portion of the image, so it is very likely that any weapons that can be correctly detected at an IoU of 0.5



(a) IoU = 0.5 (b) IoU = 0.9

(c) IoU = 0.5 (d) IoU = 0.9

Figure 4.28: True detection vs false alarm rates for handguns (top row) and long guns (bottom row) using 0.5 and 0.9 IoU thresholds.

will still be detected at a much higher threshold. This is due to the margin for error when predicting bounding boxes for small objects being very low compared to larger objects. Finally, obtaining a low false alarm rate is critical in weapon detection systems because these systems will often be monitored by a human who may ignore weapon alarms if false detections become common. While the current false alarm rate is acceptable for this work, further improvement is needed in order to operate the model in a real-world scenario (at a full frame rate). To visualize how challenging this dataset is, figures 4.29 and 4.30 show examples of successful and unsuccessful firearm detection.

Figure 4.29: Examples of successful (top) and unsuccessful (bottom) detection of long guns.

Figure 4.30: Examples of successful (top) and unsuccessful (bottom) detection of hand guns.

Chapter 5

Conclusion

5.1 Acquisition and Preprocessing

5.1.1 Facial Attribute Analysis: Mugshot Data

I investigated the advantages of rapid categorization of factors that impact face recognition performance when processing large scale face datasets collected under constrained and unconstrained conditions. To perform the experiments I used three databases, Good Quality Face Dataset, Multiple Encounter Dataset II, and a combination of the two. I proposed a software toolkit that uses multiple trained classifiers to classify face images as frontal or non-frontal, eyes open or closed, and presence or absence of glasses. To perform this classification I trained a variety of algorithms with 10-fold cross-validation, including SVMs using LBP and HOG features as well as two convolutional neural networks.

Several scenarios were trained for each factor, testing 23 different conventional models with a number of kernel functions in order to select the model and kernel function combination that best classified each factor. CNNs were optimized to find the ideal parameters for training and testing. Experimental results show that the models were able to classify each factor in the most challenging database at least 90% of the time, and over 99% for all factors in DB3 after implementing score level fusion. The most challenging factor was frontal and non-frontal faces from DB2. This is likely due to the subjective nature of the labeling process of these face images as either frontal or non-frontal because many of them were very close to being in either class. The same can be said of the eyes open or closed data, where a majority of the misclassified samples were eyes that were very slightly open. When combined with variations in expression, lighting, distance, and background, many of these samples proved to be quite challenging to classify correctly.

Based on the results I conclude that a toolkit which almost simultaneously classifies several well-known factors that affect facial recognition systems can be very beneficial to law enforcement and forensic operators at identifying individuals in the gallery. The use of hand crafted features with well-known models such as SVMs and popular CNNs can quickly find and categorize well over 90% of face images in a large database correctly, raising the overall quality of images to match against the gallery by excluding or grouping poor quality faces, or even enhancing them. Future improvement for this work could include additional factors that affect FR performance.

5.1.2 Facial Attribute Analysis: Cellphone Data

In this work I investigated the strengths and weaknesses of multiple binary facial attribute classifiers on diverse datasets with both conventional and deep learning models. I used classifiers previously trained on a database that had no images captured with a mobile device and assessed them on a database with scenarios that were both quite similar (indoors) and very different (outdoors) using an iPhone 5S. By retraining on only the mobile phone database and a combination of the two databases, I was able to dramatically improve classification performance and identify weaknesses. While frontal face and glasses classification were well over 95% accurate in many scenarios, the eyes open or closed classifiers were nearly 10% worse. Most of these misclassifications were due to the extreme illumination conditions found on many of the outdoor samples that resulted in a significant number of the subjects squinting their eyes into a barely open position. Despite these weaknesses, fast and accurate facial attribute estimation on both traditional camera and mobile phone images can be an essential step for the improvement of many areas of research, including face recognition and continuous active authentication on mobile devices. For future work, these methods could be scaled to include additional data captured from other mobile devices to asses the impact of different mobile sensors. I could also collect and label more data and train new classifiers to include any number of additional facial attributes that can help improve face recognition. Finally, improvements to the eye attribute is needed. This attribute would benefit from several techniques, including further image processing steps to correct illumination and utilizing fusion with multiple classifiers.

5.1.3 Facial Attribute Analysis: Masked Data

In this work, I assessed nine CNN architectures for classification of face mask compliance. A dataset of 100 subjects was collected and annotated for this task.

All subjects were wearing a mask in a compliant state with the mask covering their nose and mouth, and a non-compliant state with no mask on. Additionally, the data was augmented with synthetic masks where the mask is either sitting below the nose or below the mouth to account for common instances when a mask is worn but not in compliance with the current CDC mask wearing guidelines. The synthetic masks that were applied to visible images differ in shape, color, and pattern to introduce as much variance as possible into the dataset.

After assessing all classification models, I observed that the thermal band offers a more accurate option for mask compliance classification, classifying 100% of all faces correctly except for the AlexNet classifier. The models trained on the visible data are nearly as good, with accuracy well over 99% for all models. Additionally, the SqueezeNet visible model, even though it was trained on a small high-quality dataset, was able to achieve accuracy on the FMLD test set that is only 9% less accurate than the same model that was trained on FMLD in Batagelj et al., 2021.

Future work could involve the creation of a larger and more challenging dataset, especially the thermal portion. This may include more subjects with off-pose samples instead of only full-frontal samples. Additionally, data captured outdoors and at longer distances would also benefit future research, as well as assessing face detection and recognition with various levels of mask compliance.

5.1.4 MultiSpectral Face Dataset Collection

The MILAB-VTF(B) dataset is a multi-distance, unsynchronized, paired thermal-visible face dataset that was collected using the latest MWIR imaging sensors. The dataset is diverse with respect to ethnicity, age, and gender. It consists of 400 identities that were separated into training and evaluation sets that follow standard face verification protocols. Algorithmically and manually annotated face bounding boxes and keypoints are available for evaluation. In the future, a curated version will become publicly available, assisting the research community by further closing the gap of MWIR-Visible datasets availability. .

5.2 Keypoint Detection

5.2.1 Facial Landmark Detection

In this work, I presented a thermal-to-visible face synthesis pipeline for accurate detection of facial landmarks. An extensive evaluation of two common

synthesis methods, latent-guided and reference-guided, was conducted using StarGAN v2 to identify the preferred approach for future synthesis experiments. I also implemented the CUT synthesis method for further comparison. Landmark detection results on the synthesized data were reported using the HRNet facial landmark detector, previously trained on visible face data. I also compared the synthesis performance against HRNet models that were fine-tuned on thermal and multispectral datasets. I found that synthesis-based approaches can work very well and yield satisfactory results. However, recently developed synthesis models such as StarGAN v2 are still not able to outperform the traditional technique of fine-tuning previously trained visible models on thermal data, provided that sufficient amounts of data exist. The results on the ARL-VTF baseline dataset showed that the lowest synthesis model error achieved, i.e., 0.0391, is still considerably higher than the HRNet model trained solely on thermal data that yielded and overall error of 0.0186, an improvement of 110%. The fine-tuned models also achieved state-of-the-art landmark detection results on the ARL-VTF dataset when compared to other methods discussed in the open literature, and a new baseline for the MILAB-VTF(B) dataset at all distances.

An important point for consideration moving forward is that the latent-guided and reference-guided synthesis methods are suitable for different use cases. Reference-guided synthesis produced the best results, but for real-world applications this method has fewer use cases, as a reference image that is the same or similar to the face that you wish to synthesize may be required. This is especially true when there will be additional processes after landmark detection, such as face recognition. Latent-guided synthesis, however, is a very realistic use case because it synthesizes any face without the requirement of a reference image. Future work could include methods to improve the face synthesis process for the purpose of facial landmark detection and face verification using the ARL-VTF and MILAB-VTF(B) datasets.

5.3 Feature Extraction and Matching

5.3.1 Race and Gender Classification for Cross-spectral Face Recognition

In this work I investigated race and gender classification of MWIR images for thermal-to-visible reference-guided cross-spectral face recognition. I trained two networks for race and gender classification. The first was EfficientNet, an accurate and efficient network that reports state-of-the-art results on common benchmark datasets, and a multitask network based on the VGG architecture.

Both networks reported strong performance in terms of precision, recall, and F1 scores for gender classification, with F1 scores from the multitask network nearing 99%. Race classification was much more challenging, especially the Asian class where results never surpassed 50% for any of the reported metrics. The White class was the easiest to classify, with a high F1 score of 83%. Face recognition results were improved when using the race and gender classifiers I trained to select the reference image that is used during the synthesis process when compared to selecting a random race or gender. The identification scenario proved to be more challenging that matching in the verification scenario for all experiments. When using the trained classifiers for guiding the synthesis process, a high AUC score of 69.24% was reported using the EfficientNet model and VGG-Face, while a the best rank-1 and rank-10 identification performance was observed using the multitask network and VGG-Face with scores of 7.25% and 39.13% respectively. It is clear that verification is the only viable method of cross-spectral face recognition using synthesized face images at present, as seen by the poor performance in the identification experiments. The challenging nature of the identification task in cross-spectral face recognition is also likely why I could find no other papers reporting identification results in similar works. I believe that reporting these initial baseline identification results on the MILAB-VTF(B) dataset is an important starting point for future research and will encourage others to report their identification results that improve upon the results presented here.

5.4 Other Applications

5.4.1 Firearm Detection

In this work I evaluated which data augmentation techniques improve or decrease object detection performance on a novel firearm database using only real-world surveillance footage. I detected two classes of firearms, namely handguns and long guns. I used the same Faster R-CNN ResNet-50 architecture for all experiments and explored several augmentation techniques. Next, I found optimal learning rates and optimization algorithms to create an accurate object detection model for surveillance footage captured in a wide array of scenarios and resolutions. The experiments show detection performance improvement in all reported metrics by as much as 13% in some scenarios when compared to the baseline. The trained model also remains very accurate as the IoU threshold is increased up to a threshold of 0.9 and is robust to many challenging factors, including illumination and pose.

Although the proposed method works well, there are limitations with the database and training methods. Most importantly, the results and selected augmentations are likely very dependent upon the database. The model will likely perform very poorly on data not captured in a surveillance or CCTV setting. Next, more videos are needed to increase the variety of weapons and scenarios in the database. Challenging factors like occlusion, low resolution, and pose require additional data for more accurate detections. More data will also lead to better training methods and allow for train and test sets to be split by video instead of the current method, in which I sample images randomly from each video to create train and test sets. Having a test set with samples from videos that have not been trained on will result in a better understanding of how well the detection system performs on new data. Additionally, since someone using a firearm is a rare occurrence, assessment of how many false positives are triggered by the model in surveillance videos with no weapons present will provide a better understanding of model precision.

Future work could include assessing several different state-of-the-art architectures, including YOLO-based, SSD, and other recently developed detection models. Augmentation could be further explored using generative models. An analysis of selecting bounding box anchors prior to training is also an important factor to explore. Finally, an assessment of how video resolution affects performance, especially on the handgun class where instances are often very small, would be very valuable for future work.

Bibliography

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

Abbasi, S., Abdi, H., & Ahmadi, A. (2021). A face-mask detection approach based on yolo applied for a new collected dataset. *2021 26th International Computer Conference, Computer Society of Iran (CSICC)*, 1–6.

Alorf, A., & Abbott, A. L. (2017). In defense of low-level structural features and SVMs for facial attribute classification: Application to detection of eye state, mouth state, and eyeglasses in the wild. *Biometrics (IJCB), 2017 IEEE International Joint Conference on*, 599–607.

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

Anwar, A., & Raychowdhury, A. (2020). Masked face recognition for secure authentication. *arXiv preprint arXiv:2008.11104*.

Basbrain, A. M., Al-Taie, I., Azeez, N., Gan, J. Q., & Clark, A. (2017). Shallow convolutional neural network for eyeglasses detection in facial images. *Computer Science and Electronic Engineering (CEEC), 2017*, 157–161.

Batagelj, B., Peer, P., Štruc, V., & Dobrišek, S. (2021). How to correctly detect face-masks for covid-19 from visual information? *Applied Sciences*, *11*(5), 2070.

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., & Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, *35*(12), 2930–2940.

Bourdev, L., Maji, S., & Malik, J. (2011). Describing people: A poselet-based approach to attribute classification. *2011 International Conference on Computer Vision*, 1543–1550.

Bourlai, T. (2016). *Face recognition across the imaging spectrum*. Springer.

Bourlai, T., & Hornak, L. A. (2016). Face recognition outside the visible spectrum. *Image and Vision Computing*, *55*, 14–17.

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

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

Bourlai, T., Pryor, R. R., Suyama, J., Reis, S. E., & Hostler, D. (2012). Use of thermal imagery for estimation of core body temperature during precooling, exertion, and recovery in wildland firefighter protective clothing. *Prehospital Emergency Care*, *16*(3), 390–399.

Bourlai, T., Ross, A., Chen, C., & Hornak, L. (2012). A study on using midwave infrared images for face recognition. *Sensing Technologies for Global Health, Military Medicine, Disaster Response, and Environmental Monitoring II; and Biometric Technology for Human Identification IX*, *8371*, 8371IK.

Boutros, F., Damer, N., Kolf, J. N., Raja, K., Kirchbuchner, F., Ramachandra, R., Kuijper, A., Fang, P., Zhang, C., Wang, F., et al. (2021). MFR 2021: Masked face recognition competition. *2021 IEEE International Joint Conference on Biometrics (IJCB)*, 1–10.

Breiman, L. (2017). *Classification and regression trees*. Routledge.

Burgos-Artizzu, X. P., Perona, P., & Dollár, P. (2013). Robust face landmark estimation under occlusion. *Proceedings of the IEEE international conference on computer vision*, 1513–1520.

Cao, J., Li, Y., & Zhang, Z. (2018). Partially shared multi-task convolutional neural network with local constraint for face attribute learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4290–4299.

Champod, C., & Tistarelli, M. (2017). Biometric technologies for forensic science and policing: State of the art. In *Handbook of biometrics for forensic science* (pp. 1–15). Springer.

Chavda, A., Dsouza, J., Badgujar, S., & Damani, A. (2021). Multi-stage cnn architecture for face mask detection. *2021 6th International Conference for Convergence in Technology (I2CT)*, 1–8.

Chen, C., & Ross, A. (2011). Evaluation of gender classification methods on thermal and near-infrared face images. *2011 International Joint Conference on Biometrics (IJCB)*, 1–8.

Chen, C., & Ross, A. (2019). Matching thermal to visible face images using a semantic-guided generative adversarial network. *2019 14th IEEE In-*

ternational Conference on Automatic Face & Gesture Recognition (FG 2019), 1–8.

Chen, Q., Huang, J., Feris, R., Brown, L. M., Dong, J., & Yan, S. (2015). Deep domain adaptation for describing people based on fine-grained clothing attributes. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5315–5324.

Choi, Y., Uh, Y., Yoo, J., & Ha, J.-W. (2020). Stargan v2: Diverse image synthesis for multiple domains. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8188–8197.

Chowdary, G. J., Punn, N. S., Sonbhadra, S. K., & Agarwal, S. (2020). Face mask detection using transfer learning of inceptionv3. *International Conference on Big Data Analytics*, 81–90.

Chu, W.-T., & Liu, Y.-H. (2019). Thermal facial landmark detection by deep multi-task learning. *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, 1–6.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Crawshaw, M. (2020). Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, *1*, 886–893.

Damer, N., Boutros, F., Süßmilch, M., Fang, M., Kirchbuchner, F., & Kuijper, A. (2021). Masked face recognition: Human vs. machine. *arXiv preprint arXiv:2103.01924*.

Damer, N., Grebe, J. H., Chen, C., Boutros, F., Kirchbuchner, F., & Kuijper, A. (2020). The effect of wearing a mask on face recognition performance: An exploratory study. *2020 International Conference of the Biometrics Special Interest Group (BIOSIG)*, 1–6.

Dantcheva, A., Elia, P., & Ross, A. (2015). What else does your biometric data reveal? a survey on soft biometrics. *IEEE Transactions on Information Forensics and Security*, *11*(3), 441–467.

Dantcheva, A., Velardo, C., D'angelo, A., & Dugelay, J.-L. (2011). Bag of soft biometrics for person identification: New trends and challenges. *Multimedia Tools and Applications*, *51*, 739–777.

Dargan, S., & Kumar, M. (2020). A comprehensive survey on the biometric recognition systems based on physiological and behavioral modalities. *Expert Systems with Applications*, *143*, 113114.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

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

Deng, W., Hu, J., Wu, Z., & Guo, J. (2017). Lighting-aware face frontalization for unconstrained face recognition. *Pattern Recognition*, *68*, 260–271.

Dessimoz, D., & Champod, C. (2008). Linkages between biometrics and forensic science. In *Handbook of biometrics* (pp. 425–459). Springer.

Di, X., Riggan, B. S., Hu, S., Short, N. J., & Patel, V. M. (2019). Polarimetric thermal to visible face verification via self-attention guided synthesis. *2019 International Conference on Biometrics (ICB)*, 1–8.

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

Di, X., Zhang, H., & Patel, V. M. (2018). Polarimetric thermal to visible face verification via attribute preserved synthesis. *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 1–10.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *International workshop on multiple classifier systems*, 1–15.

Ding, C., Xu, C., & Tao, D. (2015). Multi-task pose-invariant face recognition. *IEEE Transactions on Image Processing*, *24*(3), 980–993.

Ding, X., & Wang, L. (2011). Facial landmark localization. In *Handbook of face recognition* (pp. 305–322). Springer.

Dong, Q., Gong, S., & Zhu, X. (2017). Class rectification hard mining for imbalanced deep learning. *Proceedings of the IEEE International Conference on Computer Vision*, 1851–1860.

Dong, Y., Zhang, Y., Yue, J., & Hu, Z. (2016). Comparison of random forest, random ferns and support vector machine for eye state classification. *Multimedia Tools and Applications*, *75*(19), 11763–11783.

Du, C., & Su, G. (2005). Eyeglasses removal from facial images. *Pattern Recognition Letters*, *26*(14), 2215–2220.

Dutta, A., Günther, M., El Shafey, L., Marcel, S., Veldhuis, R., & Spreeuwers, L. (2015). Impact of eye detection error on face recognition performance. *IET biometrics*, *4*(3), 137–150.

Eddine, B. D., Dos Santos, F. N., Boulebtateche, B., & Bensaoula, S. (2018). Eyelsd a robust approach for eye localization and state detection. *Journal of Signal Processing Systems*, *90*(1), 99–125.

Egiazarov, A., Mavroeidis, V., Zennaro, F. M., & Vishi, K. (2020). Firearm detection and segmentation using an ensemble of semantic neural networks.

El-Sayed, M. A., & Khafagy, M. A. (2014). An identification system using eye detection based on wavelets and neural networks. *1401.5108*.

FBI. (1999). *Active shooter incidents in the united states in 2018*. Retrieved April 1, 2020, from https://www.fbi.gov/file-repository/active-shooter-incidents-in-the-us-2018-041019.pdf/view

Fernandez-Carrobles, M. M., Deniz, O., & Maroto, F. (2019). Gun and knife detection based on faster r-cnn for video surveillance. *Iberian Conference on Pattern Recognition and Image Analysis*, 441–452.

Fix, E., & Hodges Jr, J. L. (1951). *Discriminatory analysis-nonparametric discrimination: Consistency properties* (tech. rep.). California Univ Berkeley.

Fogelton, A., & Benesova, W. (2016). Eye blink detection based on motion vectors analysis. *Computer Vision and Image Understanding*, *148*, 23–33.

Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). Gan-based data augmentation for improved liver lesion classification.

Fu, C., Wu, X., Hu, Y., Huang, H., & He, R. (2021). Dvg-face: Dual variational generation for heterogeneous face recognition. *IEEE transactions on pattern analysis and machine intelligence*, *44*(6), 2938–2952.

Ge, S., Li, J., Ye, Q., & Luo, Z. (2017). Detecting masked faces in the wild with LLE-CNNs. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2682–2690.

Gelana, F., & Yadav, A. (2019). Firearm detection from surveillance cameras using image processing and machine learning techniques. In *Smart innovations in communication and computational sciences* (pp. 25–34). Springer.

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

Gonog, L., & Zhou, Y. (2019). A review: Generative adversarial networks. *2019 14th IEEE conference on industrial electronics and applications (ICIEA)*, 505–510.

González-Ortega, D., Díaz-Pernas, F., Antón-Rodríguez, M., Martínez-Zarzuela, M., & Díez-Higuera, J. (2013). Real-time vision-based eye state detection for driver alertness monitoring. *Pattern Analysis and Applications*, *16*(3), 285–306.

Gonzalez-Sosa, E., Fierrez, J., Vera-Rodriguez, R., & Alonso-Fernandez, F. (2018). Facial soft biometrics for recognition in the wild: Recent works, annotation, and cots evaluation. *IEEE Transactions on Information Forensics and Security*, *13*(8), 2001–2014.

Goodfellow, I., Jean, P.-A., Mehdi, M., Bing, X., David, W.-F., Sherjil, O., & Courville, A. (2014). Generative adversarial nets. *Proceedings of the 27th international conference on neural information processing systems*, *2*, 2672–2680.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144.

Grega, M., Łach, S., & Sieradzki, R. (2013). Automated recognition of firearms in surveillance video. *2013 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, 45–50.

Grega, M., Matiolański, A., Guzik, P., & Leszczuk, M. (2016). Automated detection of firearms and knives in a cctv image. *Sensors*, *16*(1), 47.

Gui, J., Sun, Z., Wen, Y., Tao, D., & Ye, J. (2021). A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 1–1. https://doi.org/10.1109/TKDE.2021.3130191

Guo, G., & Zhang, N. (2019). A survey on deep learning based face recognition. *Computer vision and image understanding*, *189*, 102805.

Halima, N. B., & Hosam, O. (2016). Bag of words based surveillance system using support vector machines. *International Journal of Security and Its Applications*, *10*(4), 331–346.

Hassan, H., Yaacob, S., Radman, A., & Suandi, S. A. (2016). Eye state detection for driver inattention based on lucas kanade optical flow algorithm. *Intelligent and Advanced Systems (ICIAS), 2016 6th International Conference on*, 1–6.

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

He, K., Fu, Y., Zhang, W., Wang, C., Jiang, Y.-G., Huang, F., & Xue, X. (2018). Harnessing synthesized abstraction images to improve facial attribute recognition. *IJCAI*, 733–740.

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

Ho, H. T., & Chellappa, R. (2013). Pose-invariant face recognition using Markov random fields. *IEEE transactions on image processing*, *22*(4), 1573–1584.

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

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

Huang, C., Li, Y., Change Loy, C., & Tang, X. (2016). Learning deep representation for imbalanced classification. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5375–5384.

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

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

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

Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., & Keutzer, K. (2016). SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and $<$ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*.

Iqbal, J., Munir, M. A., Mahmood, A., Ali, A. R., & Ali, M. (2019). Orientation aware object detection with application to firearms.

Iqbal, J., Munir, M. A., Mahmood, A., Ali, A. R., & Ali, M. (2021). Leveraging orientation for weakly supervised object detection with application to firearm localization. *Neurocomputing*, *440*, 310–320.

Iranmanesh, S. M., & Nasrabadi, N. M. (2019). Attribute-guided deep polarimetric thermal-to-visible face recognition. *2019 International Conference on Biometrics (ICB)*, 1–8.

Iranmanesh, S. M., Riggan, B., Hu, S., & Nasrabadi, N. M. (2020). Coupled generative adversarial network for heterogeneous face recognition. *Image and Vision Computing*, *94*, 103861.

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

Jain, A. K., Dass, S. C., & Nandakumar, K. (2004). Can soft biometric traits assist user recognition? *Biometric technology for human identification*, *5404*, 561–572.

Jain, A. K., Klare, B., & Park, U. (2011). Face recognition: Some challenges in forensics. *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 726–733.

Jain, A. K., Klare, B., & Park, U. (2012). Face matching and retrieval in forensics applications. *IEEE multimedia*, *19*(1), 20.

Jain, A. K., & Li, S. Z. (2011). *Handbook of face recognition*. Springer.

Jain, A. K., & Ross, A. (2015). Bridging the gap: From biometrics to forensics. *Phil. Trans. R. Soc. B*, *370*(1674), 20140254.

Jalil, A. J., & Reda, N. M. (2022). Infrared thermal image gender classifier based on the deep resnet model. *Advances in Human-Computer Interaction*, *2022*.

Ji, Y., Wang, S., Lu, Y., Wei, J., & Zhao, Y. (2018). Eye and mouth state detection algorithm based on contour feature extraction. *Journal of Electronic Imaging*, *27*(5), 051205.

Jiang, M., Fan, X., & Yan, H. (2020). Retina facemask: A face mask detector. *arXiv preprint arXiv:2005.03950*, *2*.

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

Kalayeh, M. M., Gong, B., & Shah, M. (2017). Improving facial attribute prediction using semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6942–6950.

Keong, J., Dong, X., Jin, Z., Mallat, K., & Dugelay, J.-L. (2020). Multi-spectral facial landmark detection. *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–6.

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

Khabarlak, K., & Koriashkina, L. (2021). Fast facial landmark detection and applications: A survey. *arXiv preprint arXiv:2101.10808*.

Khandelwal, P., Khandelwal, A., Agarwal, S., Thomas, D., Xavier, N., & Raghuraman, A. (2020). Using computer vision to enhance safety of workforce in manufacturing in a post covid world. *arXiv preprint arXiv:2005.05287*.

Kim, K. W., Hong, H. G., Nam, G. P., & Park, K. R. (2017). A study of deep CNN-based classification of open and closed eyes using a visible light camera sensor. *Sensors*, *17*(7), 1534.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, *10*, 1755–1758.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization.

Koestinger, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2011). Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, 2144–2151.

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

Kopaczka, M., Kolk, R., Schock, J., Burkhard, F., & Merhof, D. (2018). A thermal infrared face database with facial landmarks and emotion labels. *IEEE Transactions on Instrumentation and Measurement*, *68*(5), 1389–1401.

Kopaczka, M., Schock, J., & Merhof, D. (2019). Super-realtime facial landmark detection and shape fitting by deep regression of shape model parameters. *arXiv preprint arXiv:1902.03459*.

Kowalski, M., Naruniec, J., & Trzcinski, T. (2017). Deep alignment network: A convolutional neural network for robust face alignment. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 88–97.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1097–1105.

Kumar, N., Berg, A. C., Belhumeur, P. N., & Nayar, S. K. (2009). Attribute and simile classifiers for face verification. *2009 IEEE 12th International Conference on Computer Vision*, 365–372.

Kuzdeuov, A., Koishigarina, D., Aubakirova, D., Abushakimova, S., & Varol, H. A. (2022). Sf-tl54: A thermal facial landmark dataset with visual pairs. *2022 IEEE/SICE International Symposium on System Integration (SII)*, 748–753.

Lai, J., & Maples, S. (2017). Developing a real-time gun detection classifier [Date accessed: 2019-01-10]. http://cs231n.stanford.edu/reports/2017/pdfs/716.pdf.

Lazarus, M. Z., & Gupta, S. (2016). A low rank model based improved eye detection under spectacles. *Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), IEEE Annual*, 1–6.

Le, V., Brandt, J., Lin, Z., Bourdev, L., & Huang, T. S. (2012). Interactive facial feature localization. *European conference on computer vision*, 679–692.

Lee, J., Jain, A., Tong, W., et al. (2012). Image retrieval in forensics: Tattoo image database application. *IEEE MultiMedia*, *19*(1), 40–49.

LIANG, M., XUE, Y., XUE, K., & YANG, A. (2017). Deep convolution neural networks for automatic eyeglasses removal. *DEStech Transactions on Computer Science and Engineering*, (aiea).

Lim, J., Al Jobayer, M. I., Baskaran, V. M., Lim, J. M., Wong, K., & See, J. (2019). Gun detection in surveillance videos using deep neural networks. *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1998–2002.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *European conference on computer vision*, 740–755.

Liu, L., Ouyang, W., Wang, X., Fieguth, P., Chen, J., Liu, X., & Pietikäinen, M. (2020). Deep learning for generic object detection: A survey. *International journal of computer vision*, *128*, 261–318.

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

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015b). Deep learning face attributes in the wild. *Proceedings of the IEEE international conference on computer vision*, 3730–3738.

Loey, M., Manogaran, G., Taha, M. H. N., & Khalifa, N. E. M. (2021a). Fighting against covid-19: A novel deep learning model based on yolo-v2 with resnet-50 for medical face mask detection. *Sustainable cities and society*, *65*, 102600.

Loey, M., Manogaran, G., Taha, M. H. N., & Khalifa, N. E. M. (2021b). A hybrid deep transfer learning model with machine learning methods for face mask detection in the era of the covid-19 pandemic. *Measurement*, *167*, 108288.

Lu, Y., Kumar, A., Zhai, S., Cheng, Y., Javidi, T., & Feris, R. (2017). Fully-adaptive feature sharing in multi-task networks with applications in

person attribute classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5334–5343.

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

Mallat, K., & Dugelay, J.-L. (2018b). A benchmark database of visible and thermal paired face images across multiple variations. *International Conference of the Biometrics Special Interest Group, BIOSIG 2018, Darmstadt, Germany, September*, 199–206.

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

Mandal, B., Li, L., Wang, G. S., & Lin, J. (2017). Towards detection of bus driver fatigue based on robust visual analysis of eye state. *IEEE Transactions on Intelligent Transportation Systems*, *18*(3), 545–557.

Martin, M., Xu, X., & Lane, T. B. (2016). A multimedia application for location-based semantic retrieval of tattoos. *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 1–8.

Masi, I., Chang, F.-J., Choi, J., Harel, S., Kim, J., Kim, K., Leksut, J., Rawls, S., Wu, Y., Hassner, T., et al. (2018). Learning pose-aware models for pose-invariant face recognition in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Meuwly, D., & Veldhuis, R. (2012). Forensic biometrics: From two communities to one discipline. *Biometrics Special Interest Group (BIOSIG), 2012 BIOSIG-Proceedings of the International Conference of the*, 1–12.

Mohammad, A. S., Rattani, A., & Derahkshani, R. (2017). Eyeglasses detection based on learning and non-learning based classification schemes. *Technologies for Homeland Security (HST), 2017 IEEE International Symposium on*, 1–5.

Mohan, P., Paul, A. J., & Chirania, A. (2021). A tiny CNN architecture for medical face mask detection for resource-constrained endpoints. In *Innovations in electrical and electronic engineering* (pp. 657–670). Springer.

Mokalla, S. R., & Bourlai, T. (2019). On designing MWIR and visible band based deepface detection models. *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 1140–1147.

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

Narang, N., & Bourlai, T. (2016). Gender and ethnicity classification using deep learning in heterogeneous face recognition. *2016 International Conference on Biometrics (ICB)*, 1–8.

Narang, N., Martin, M., Metaxas, D., & Bourlai, T. (2017). Learning deep features for hierarchical classification of mobile phone face datasets in heterogeneous environments. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 186–193.

Nascimento, J. C., Abrantes, A. J., & Marques, J. S. (1999). An algorithm for centroid-based tracking of moving objects. *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, *6*, 3305–3308.

Ngan, M., Grother, P., & Hanaoka, K. (2020). Ongoing face recognition vendor test (frvt) part 6a: Face recognition accuracy with masks using pre-covid-19 algorithms. https://doi.org/https://doi.org/10.6028/NIST.IR.8311

Niinuma, K., Park, U., & Jain, A. K. (2010). Soft biometric traits for continuous user authentication. *IEEE Transactions on information forensics and security*, *5*(4), 771–780.

Oh, B.-S., Toh, K.-A., Teoh, A. B. J., & Lin, Z. (2018). An analytic gabor feedforward network for single-sample and pose-invariant face recognition. *IEEE Transactions on Image Processing*, *27*(6), 2791–2805.

Ojala, T., Pietikäinen, M., & Harwood, D. (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, *29*(1), 51–59.

Olmos, R., Tabik, S., & Herrera, F. (2018). Automatic handgun detection alarm in videos using deep learning. *Neurocomputing*, *275*, 66–72.

Osia, N., & Bourlai, T. (2012). Holistic and partial face recognition in the mwir band using manual and automatic detection of face-based features. *2012 IEEE Conference on Technologies for Homeland Security (HST)*, 273–279.

Osia, N., & Bourlai, T. (2017a). Bridging the spectral gap using image synthesis: A study on matching visible to passive infrared face images. *Machine Vision and Applications*, *28*(5), 649–663.

Osia, N., & Bourlai, T. (2017b). On matching visible to passive infrared face images using image synthesis & denoising. *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, 904–911.

Osia, N., Bourlai, T., & Hornak, L. (2018). Facial surveillance and recognition in the passive infrared bands. *Surveillance in Action: Technologies for Civilian, Military and Cyber Surveillance*, 127–145.

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

Park, T., Efros, A. A., Zhang, R., & Zhu, J.-Y. (2020). Contrastive learning for unpaired image-to-image translation. *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, 319–345.

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

Pauly, L., & Sankar, D. (2015). Detection of drowsiness based on hog features and svm classifiers. *Research in Computational Intelligence and Communication Networks (ICRCICN), 2015 IEEE International Conference on*, 181–186.

Peri, N., Gleason, J., Castillo, C. D., Bourlai, T., Patel, V. M., & Chellappa, R. (2021a). A synthesis-based approach for thermal-to-visible face verification. *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, 01–08.

Peri, N., Gleason, J., Castillo, C. D., Bourlai, T., Patel, V. M., & Chellappa, R. (2021b). A synthesis-based approach for thermal-to-visible face verification. https://doi.org/10.48550/ARXIV.2108.09558

Phillips, P. J., Grother, P., & Micheals, R. (2011). Evaluation methods in face recognition. *Handbook of face recognition*, 551–574.

Poster, D., Hu, S., Nasrabadi, N., & Riggan, B. (2019). An examination of deep-learning based landmark detection methods on thermal face imagery. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 980–987. https://doi.org/10.1109/CVPRW.2019.00129

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

Poster, D. D., Hu, S., Short, N. J., Riggan, B. S., & Nasrabadi, N. M. (2021). Visible-to-thermal transfer learning for facial landmark detection. *IEEE Access*, *9*, 52759–52772.

Qian, N. (1999). On the momentum term in gradient descent learning algorithms. *Neural networks*, *12*(1), 145–151.

Qin, B., & Li, D. (2020). Identifying facemask-wearing condition using image super-resolution with classification network to prevent covid-19. *Sensors*, *20*(18), 5236.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks.

Radlak, K., & Smolka, B. (2012). A novel approach to the eye movement analysis using a high speed camera. *Advances in Computational Tools for Engineering Applications (ACTEA), 2012 2nd International Conference on*, 145–150.

Raghavendra, R., Vetrekar, N., Raja, K. B., Gad, R. S., & Busch, C. (2018). Robust gender classification using extended multi-spectral imaging by exploring the spectral angle mapper. *2018 IEEE 4th International Conference on identity, security, and behavior analysis (ISBA)*, 1–8.

Rahman, M. M., Manik, M. M. H., Islam, M. M., Mahmud, S., & Kim, J.-H. (2020). An automated system to limit covid-19 using facial mask detection in smart city network. *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 1–5.

Redmon, J., & Farhadi, A. (2017). Yolo9000: Better, faster, stronger. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7263–7271.

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 91–99.

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

Romero, D., & Salamea, C. (2019). Convolutional models for the detection of firearms in surveillance videos. *Applied Sciences*, *9*(15), 2965.

Rose, J., & Bourlai, T. (2020). On designing a forensic toolkit for rapid detection of factors that impact face recognition performance when processing large scale face datasets. *Securing Social Identity in Mobile Platforms: Technologies for Security, Privacy and Identity Management*, 61–76.

Roth, K., Lucchi, A., Nowozin, S., & Hofmann, T. (2017). Stabilizing training of generative adversarial networks through regularization. *Advances in neural information processing systems*, *30*.

Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 397–403.

Samangouei, P., & Chellappa, R. (2016). Convolutional neural networks for attribute-based active authentication on mobile devices. *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 1–8.

Samangouei, P., Patel, V. M., & Chellappa, R. (2015). Attribute-based continuous user authentication on mobile devices. *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 1–8.

Samangouei, P., Patel, V. M., & Chellappa, R. (2017). Facial attributes for active authentication on mobile devices. *Image and Vision Computing*, *58*, 181–192.

Sarfraz, M. S., & Stiefelhagen, R. (2017). Deep perceptual mapping for cross-modal face recognition. *International Journal of Computer Vision*, *122*, 426–438.

Sethi, S., Kathuria, M., & Kaushik, T. (2021). Face mask detection using deep learning: An approach to reduce risk of coronavirus spread. *Journal of Biomedical Informatics*, *120*, 103848.

Shao, M., Zhang, Y., & Fu, Y. (2018). Collaborative random faces-guided encoders for pose-invariant face representation learning. *IEEE transactions on neural networks and learning systems*, *29*(4), 1019–1032.

Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2016). Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 1528–1540.

Shen, W., & Liu, R. (2017). Learning residual images for face attribute manipulation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4030–4038.

Shijie, J., Ping, W., Peiyi, J., & Siping, H. (2017). Research on data augmentation for image classification based on convolution neural networks. *2017 Chinese automation congress (CAC)*, 4165–4170.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 60.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singleton, M., Taylor, B., Taylor, J., & Liu, Q. (2018). Gun identification using tensorflow. *International Conference on Machine Learning and Intelligent Communications*, 3–12.

Smith-Creasey, M., Albalooshi, F. A., & Rajarajan, M. (2018). Continuous face authentication scheme for mobile devices with tracking and liveness detection. *Microprocessors and Microsystems*, *63*, 147–157.

Suresh, K., Palangappa, M., & Bhuvan, S. (2021). Face mask detection by using optimistic convolutional neural network. *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, 1084–1089.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1–9.

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1701–1708.

Tan, M., & Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *International conference on machine learning*, 6105–6114.

Taskiran, M., Kahraman, N., & Erdem, C. E. (2020). Face recognition: Past, present and future (a review). *Digital Signal Processing*, *106*, 102809.

Taylor, L., & Nitschke, G. (2017). Improving deep learning using generic data augmentation.

Taylor, M. (2017). The art of facial recognition [Accessed: 2018-02-18].

Tieleman, T., & Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, *4*(2), 26–31.

Tistarelli, M., Grosso, E., & Meuwly, D. (2014). Biometrics in forensic science: Challenges, lessons and new technologies. *International Workshop on Biometric Authentication*, 153–164.

Tiwari, R. K., & Verma, G. K. (2015a). A computer vision based framework for visual gun detection using harris interest point detector. *Procedia Computer Science*, *54*, 703–712.

Tiwari, R. K., & Verma, G. K. (2015b). A computer vision based framework for visual gun detection using surf. *2015 International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO)*, 1–5.

Verma, G. K., & Dhillon, A. (2017). A handheld gun detection using faster r-cnn deep learning. *Proceedings of the 7th International Conference on Computer and Communication Technology*, 84–88.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, *1*, I–I.

Wang, J., Yuan, Y., & Yu, G. (2017). Face attention network: An effective face detector for the occluded faces. *arXiv preprint arXiv:1711.07246*.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al. (2020). Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, *43*(10), 3349–3364.

Wang, L., Chen, W., Yang, W., Bi, F., & Yu, F. R. (2020). A state-of-the-art review on image synthesis with generative adversarial networks. *IEEE Access*, *8*, 63514–63537.

Wang, S., Gao, Z., He, S., He, M., & Ji, Q. (2016). Gender recognition from visible and thermal infrared facial images. *Multimedia Tools and Applications*, *75*, 8419–8442.

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

Wang, Y., Ou, X., Tu, L., & Liu, L. (2018). Effective facial obstructions removal with enhanced cycle-consistent generative adversarial networks. *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, 210–220.

Wang, Z., Chen, Z., & Wu, F. (2018). Thermal to visible facial image translation using generative adversarial networks. *IEEE Signal Processing Letters*, *25*(8), 1161–1165.

Wang, Z., Wang, G., Huang, B., Xiong, Z., Hong, Q., Wu, H., Yi, P., Jiang, K., Wang, N., Pei, Y., et al. (2020). Masked face recognition dataset and application. *arXiv preprint arXiv:2003.09093*.

White, D., Dunn, J. D., Schmid, A. C., & Kemp, R. I. (2015). Error rates in users of automatic face recognition software. *PLoS One*, *10*(10), e0139827.

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

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

Wing, B. J. (2013). The ANSI/NIST-ITL standard update for 2011 (data format for the interchange of fingerprint, facial and other biometric information). *International Journal of Biometrics*, *5*(1), 20–29.

Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., & Zhou, Q. (2018). Look at boundary: A boundary-aware face alignment algorithm. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2129–2138.

Wu, Y., & Ji, Q. (2019). Facial landmark detection: A literature survey. *International Journal of Computer Vision*, *127*(2), 115–142.

Xiao, T., Hong, J., & Ma, J. (2018). Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. *Proceedings of the European Conference on Computer Vision (ECCV)*, 168–184.

Xu, W., Shen, Y., Bergmann, N., & Hu, W. (2018). Sensor-assisted multi-view face recognition system on smart glass. *IEEE Transactions on Mobile Computing*, *17*(1), 197–210.

Yadav, P., Gupta, N., & Sharma, P. K. (2022). A comprehensive study towards high-level approaches for weapon detection using classical machine learning and deep learning methods. *Expert Systems with Applications*, 118698.

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

Yang, X., & Bourlai, T. (2018). Video-based human respiratory wavelet extraction and identity recognition. In *Surveillance in action* (pp. 51–75). Springer.

Ying, H., Da Su, G., & Chen, J. S. (2014). Automatic eyeglasses removal of frontal facial images for face recognition. *Applied Mechanics and Materials*, *571*, 697.

Zafeiriou, S., Trigeorgis, G., Chrysos, G., Deng, J., & Shen, J. (2017). The menpo facial landmark localisation challenge: A step towards the solution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 170–179.

Zhang, H., Riggan, B. S., Hu, S., Short, N. J., & Patel, V. M. (2019). Synthesis of high-quality visible faces from polarimetric thermal faces using generative adversarial networks. *International Journal of Computer Vision*, *127*(6), 845–862.

Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, *23*(10), 1499–1503.

Zhang, N., Paluri, M., Ranzato, M., Darrell, T., & Bourdev, L. (2014). Panda: Pose aligned networks for deep attribute modeling. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1637–1644.

Zhang, T., Wiliem, A., Yang, S., & Lovell, B. (2018). Tv-gan: Generative adversarial network based thermal to visible face recognition. *2018 international conference on biometrics (ICB)*, 174–181.

Zhang, X., & Gao, Y. (2009). Face recognition across pose: A review. *Pattern Recognition*, *42*(11), 2876–2896.

Zhang, X., Pham, D.-S., Venkatesh, S., Liu, W., & Phung, D. (2015). Mixed-norm sparse representation for multi view face recognition. *Pattern recognition*, *48*(9), 2935–2946.

Zhao, L., Wang, Z., Zhang, G., Qi, Y., & Wang, X. (2017). Eye state recognition based on deep integrated neural network and transfer learning. *Multimedia Tools and Applications*, 1–24.

Zhao, M., Zhang, X., Shi, Z., Chen, T., & Zhang, F. (2018). Eyeglasses detection, location and frame discriminant based on edge information projection. *Multimedia Tools and Applications*, *77*(12), 14931–14949.

Zheng, X., Guo, Y., Huang, H., Li, Y., & He, R. (2018). A survey to deep facial attribute analysis. *arXiv preprint arXiv:1812.10265*.

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

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), 43–76.

Zoph, B., Vasudevan, V., Shlens, J., & Le, Q. V. (2018). Learning transferable architectures for scalable image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8697–8710.

Zou, Z., Chen, K., Shi, Z., Guo, Y., & Ye, J. (2023). Object detection in 20 years: A survey. *Proceedings of the IEEE*.