Facial-Liveliness-Verification
for Monocular Real-Time-Systems

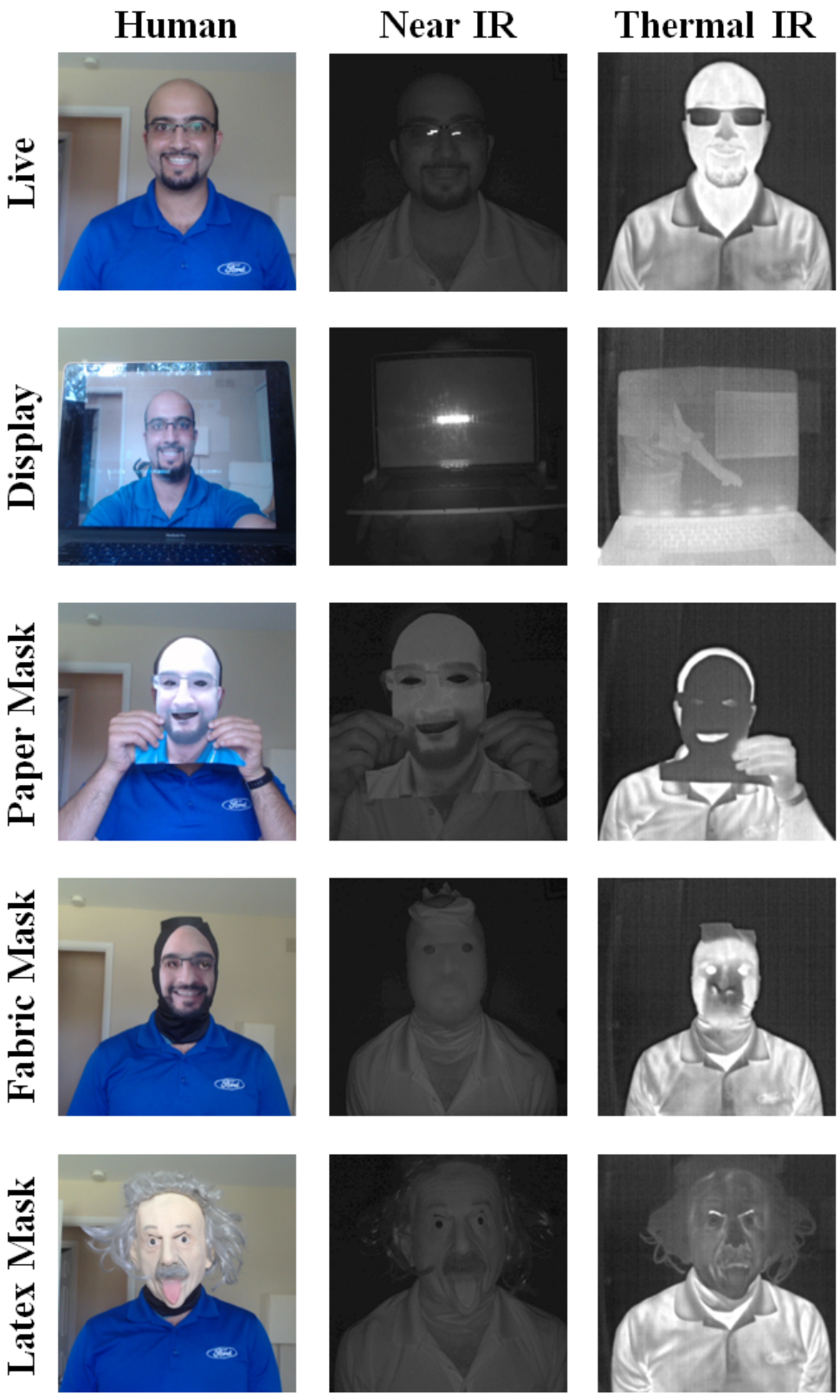
by

Ali Hassani

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical and Computer Engineering)
in the University of Michigan - Dearborn
2022

Doctoral Committee:

Professor Hafiz Malik, Chair
Assistant Professor Mohamed Abouelenien
Associate Professor Sridhar Lakshmanan
Associate Professor Samir Rawashdeh
Professor Adnan Shaout


Human
Near IR
Thermal IR
Live
Display
Paper Mask
Fabric Mask
Latex Mask

Ali Hassani
alihassa@umich.edu
ORCID iD: 0000-0003-0097-6807

DEDICATION

I dedicate this dissertation to my parents, my significant other and my grandmother. Their support carried me through all the roadblocks and pitfalls of doctoral research. I could not have completed this without them.

ACKNOWLEDGMENTS

Thank you to all the people who made this dissertation possible. Family members Mohammad Hassani, Nooshin Sobhani and Alexandra Taylor for their continuous support. Advisers Dr. Hafiz Malik (faculty) and Jon Diedrich (Ford PI) for their guidance throughout this research. Faculty members Dr. Paul Richardson and Dr. Adnan Shaout, and Dr. Dimitar Filev of the Ford Motor Company for petitioning an exception so that I could enroll as a full-time engineer. Collaborators Zaid El Shair, Rafi Ud Daula Refat, Jun Lin, Hamid Golgiri and Reates Curry for their considerable help. Ford management John Van Wiemeersch, Justin Miller and Nicholas Colella for orchestrating a corporate role that is well aligned with this research. Ford counsel Frank Lollo for negotiating the alliance grant intellectual-property agreement. UM staff Amanda Donovan, Michael Hicks and Geoffrey Hosker for their assistance in preparing the doctoral documentation. Any others who helped with this dissertation and is not mentioned, I appreciate your time and efforts as well. This research is funded by the Ford-UM Alliance Grant, Biometric Forensics.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

LIST OF ACRONYMS

ACER average-classification-error-rate

APCER attack-presentation-classification-error-rate

ANT auxiliary-noise-task

BJL biometric-joint-learning

CFA color-filter-array

CNN convolutional-neural-network

DCT discrete-cosine-transform

DL deep-learning

DWT discrete-wavelet-transform

ELA error-level-analysis

FLV facial-liveliness-verification

FPN feature-pyramid-network

FR face-recognition

FSA face-swap-attack

FZV face-zone-verification

GAN generative-adversarial-network

HOG histogram-of-oriented-gradients

HMI human-machine-interface

IIV image-integrity-verification

LBP local-binary-pattern

LWIR long-wave-infrared

MTL multi-task-learning

MS material-spectroscopy

NIR near-infrared

NIST National Institute of Standards and Technology

NPCER nominal-presentation-classification-error-rate

PCA principle-component-analysis

PCE peak-correlation-energy

PRNU photo-response-non-uniformity

PSA physical-spoof-attack

R-CNN region-based convolutional-neural-network

RCH Ranked-channel-histograms

RFS random-Fourier-series

RTS real-time-systems

SDA service-denial-attack

SoC system-on-chip

SWIR short-wave-infrared

ZEV zonal-expected-value

ABSTRACT

Face-recognition is becoming the go-to authentication method. It is convenient: simply look at the camera for instant recognition. Attackers, however, can expose vulnerabilities by "replaying" an enrolled user. The primary concern here is the physical-spoof-attack. Attackers can acquire a representative image from social media and create a realistic looking facsimile (e.g., paper-mask) for authentication. This attack is rather popular for its efficacy and simplicity; despite this, there are few reliable monocular detection methods. Alternatively, attackers can tamper the camera stream by placing an injection device. The face-swap-attack similarly presents an acquired image of the victim, this time as a photo-realistic image alteration using machine-learning. This attack is new and does not yet have a computationally efficient means of detection. The goal of this dissertation is to address both problems in a fashion that is monocular, single-frame and computationally-efficient. A series of four physics-informed facial-liveliness-verification frameworks are presented to achieve these goals. Performance evaluation shows best-in-class accuracy where all algorithms are optimized for real-time-systems. These results are discussed and concluded with proposed future works.

CHAPTER 1

Introduction

Face-recognition (FR) is becoming the go-to authentication method for digital platforms. People can simply look at a camera and be instantly recognized. This technology is made possible through deep-learning methods. State-of-the-art identification networks today are able to discern one person from over 50,000 (1). A growing problem, however, is attackers can "replay" enrolled users to spoof authentication. With just a photo from social-media, attackers can present an enrolled face through physical facsimiles or digitally injection. This dissertation addresses these vulnerabilities through a proposed series of facial-liveliness-verification (FLV) methods. The goal is to design physics-informed algorithms that can be mass deployed on monocular real-time-systems (RTS).

1.1 Face-Recognition Applications

To properly appreciate the value of attacking FR systems, it is important to understand how pervasive the applications are. FR is popular because it is arguably the most convenient way to authenticate. It not only frees users from having to memorize a password or carry a device, but also offers the greatest freedom out of traditional biometrics. Other common methods, such as fingerprint and iris recognition, require the user precisely interact with the sensor. The goal here is to be seamless, where any general look towards the camera is instantly authenticated. This combination of seamless and security results in FR being the fastest growing authentication method, projecting to achieve $8.5 billion in annual revenue by 2025 (2).

Smart-devices are arguably championing this rapid adoption (3). Recent studies show over 100 million currently-deployed smart-phones have FR, with the anticipation of up to 90% of new-phones will offer it by 2024 (4). This trend is similarly happening with laptops. Windows Hello offers a standardized FR authenticator for plug-and-play near-infrared (NIR) cameras; this is supported on any Windows device (including desktops) (5).

Other industries are now also using FR as an experience-differentiator. Commercial buildings use it as a means of building access (6). Hotels are planning on greeting guests upon entry (anticipated 70% adoption by 2024) (7). Airports are using it for passport verification (8), now required at all of the top 20 US airports (by travel volume) (9). Today even automobiles are personalizing cabins and allowing owners to function as a biometric key (10). These industry disruptions are happening globally, where China is in general leading the pack for convenience and surveillance applications (11).

This pervasive application is why there is so much value to exposing FR vulnerabilities. If an attacker can design a quality spoofing method, they can effectively take over a victim's life. They can authenticate into a victim's phone, home and even vehicle. Or conversely, can act maliciously and frame the user to surveillance systems. These types of attacks clearly need to be addressed in a mass-deployable fashion.

1.2 Technological Challenges

Despite its popularity, FR is still a relatively nascent technology. The areas of greatest development are the detection and identification algorithms. State-of-the-art methods can observe tiny faces (12) and recognize them at scale (13). For this reason, the challenges presented focus on liveliness.

Physical-spoof-attacks (PSAs) "replay" a user's facsimile to the camera (14). Attackers typically find "replay" sources from social-media and make them into facsimiles. Common examples include pictures, videos and simple-masks printed onto paper or fabric(14). Very simple presentations can be detected with basic motion methods (e.g., eye blink (15) or heart rate (16)), however, most typically requires depth sensing approaches (17; 18). These approaches are expensive and a key part of why many FR systems are still vulnerable to the PSA.

Digital attacks manipulate the camera-data. With advances in compute and AI technologies, attackers can alter the image-stream to contain a valid user. Face-swap-attacks (FSAs) are getting relatively popular as they can photo-realistically present an enrolled user and there are a plethora of available swapping-algorithms. Alternatively, attackers can also perform the service-denial-attack (SDA) by using the synthesis networks to photo-realistically remove faces. It is noted there is a rich history of identifying camera source and detecting basic tampering attacks (19; 20). The challenge here is detecting these photo-realistic methods in a fashion that is computationally efficient.

Between physical and digital spoofing attacks, it is clear FLV is an important area of research. In particular, there is a need for methods that are compatible with current RTSs (e.g., monocular and computationally efficient) to ensure mass-deployment. With that said, it is important to also consider issues that impact the user experience: false-rejections, latency and privacy. These are viewed as secondary within this security-minded research, as there are suitable available solutions.

False-rejections occur when the presented face deviates from the enrollment expectation. Environment, facial perspective, aging, weight change, injury and threshold sensitivity all can factor in mistaking a valid user (21; 22; 23). Thankfully, these can be mitigated several ways. The enrollment process can be improved by adding more perspectives and updating it periodically (24). Algorithms can incorporate additional context to be less sensitive to variations in pose or environment. If all else fails, a robust and easy to use backup should be provided (1).

Latency correlates with convenience. If the FR system is too slow, users will tend to opt for alternative authentication methods. This issue is partially mitigated by the introduction of neural system-on-chips (SoCs), which can run deep-learning networks in real-time (25). These is a hardware solution, however, and drives cost. A simple, cost-effective solution is to provide fast feedback. Users tend to be more tolerant of delays when they understand the system behavior.

Lastly, privacy is a concern whenever using camera-monitoring technologies. Users need trust that their biometric-data is secure, and that the FR system will only authenticate when intended. This is typically mitigated using a couple of techniques. Biometric-data is typically stored in a trusted environment and never transmitted in a raw format (1). To help forge trust with the system, it is customary to provide a human-machine-interface (HMI) to communicate status and camera activity. Users can be given additional control by cooperating, where they look at the camera or perform a gesture to signify intent to authenticate.

1.3 Problem Statement

The central problem statement is to **verify facial-liveliness in a monocular, single-frame fashion**. There are literally millions of FR authenticators deployed that are vulnerable to spoofing attacks. Coming up with a security approach that can implemented with just a software-update would be a tremendous contribution. The primary FR vulnerabilities are annotated in Fig. 1.1

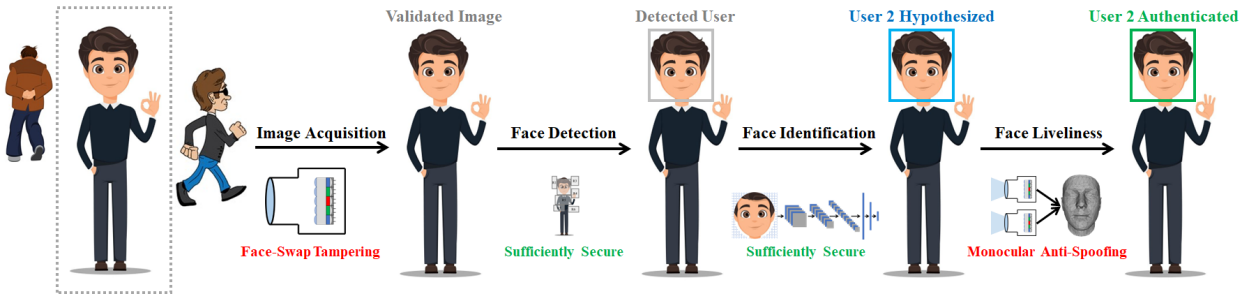


Figure 1.1: Face-recognition vulnerabilities.

1.4 Research Objectives

The research aims to address the described gaps in efficiently FLV methodology. The primary motivation is to address the PSA. As introduced in the challenges, attackers can spoof authenticators with a trivial amount of effort and cost. Despite the attack's popularity, most state-of-the-art methods require 3D sensing (18). The secondary motivation is to address photo-realistic tampering through FSA or SDA. While generally more rare (they require physical access to the sensor for tampering), current platforms are vulnerable to them (26).

This dissertations addresses both of these attack vectors using an optical-forensics framework. This underlying physics is identified and applied with robust algorithms. The contributions are as follows:

1. An end to end threat-model for face-recognition spoofing
2. A novel material-spectroscopy approach to physical-spoof detection
3. A novel noise-synthesis framework to evaluate camera and environment effects
4. A novel auxiliary-noise-task framework to maximize spectroscopy robustness
5. A novel image-integrity-verification method employing camera-noise

1.5 Dissertation Outline

In summary, this dissertation addresses FLV in a monocular, single-frame fashion. This process starts with presenting a basic FR pipeline overview in Chapter-2. Once the general algorithms are understood, an end-to-end threat-model with vulnerability analysis is presented in Chapter-3. A literature survey is then presented on the relevant state-of-the-art liveliness methods in Chapter-4. These chapters serve as the background section of the dissertation, aimed at helping the reader understand the fundamental vulnerabilities and mitigation strategies.

The contributions start with presenting the physics-informed frameworks in Chapter-5. First, the PSA is addressed. This starts with presenting a novel near-infrared material-spectroscopy approach, classifying liveliness from reflectance-patterns, in Chapter-6. Next, the spectroscopy method's robustness to camera and environmental noise is presented in Chapter-7, which includes a novel approach to generating noise synthetically. The methodology is then enhanced with a novel noise-based approach to multi-task-learning in Chapter-8. With the PSA addressed, the last FLV framework addresses photo-realistic tampering (FSA and SDA). A novel image-integrity-verification framework is presented in Chapter-9.

These chapters conclude the FLV research. All of the desired objectives are met, where the proposed algorithms meet the intended goals and perform extremely robustly. For completeness, a related application on improving face-identification robustness by encoding semantic-segmentation features is also presented in Chapter-10. The dissertation contributions are then concluded in Chapter-11, with some proposed future works. The author desires that anyone reading this dissertation to find the contributions useful, and enjoy it as much they do.

CHAPTER 2

Face-Recognition Background

Face-recognition (FR) is a camera-based authentication service. This chapter starts with introducing the traditional authentication pipeline, briefly discussing each algorithm. Next, some common applications are discussed. If further background information is desired, a glossary of relevant terminology can also be found in Appendix-A.

2.1 Authentication Pipeline

The FR authentication pipeline is a process of detecting faces and then analyzing them for identity and liveliness. The traditional approach is to acquire an image, detect the faces, determine the identity and lastly verify liveliness. This pipeline is illustrated in Fig. 2.1. Note, however, this pipeline can change order depending upon the available hardware. For example, 3D sensing methods are extremely effective at facial-liveliness-verification; in these situations liveliness can be verified first, only addressing identity for live faces. Additionally, single-shot detection and identification is an active research area (27). There is computational value to combining these algorithms onto a single network.

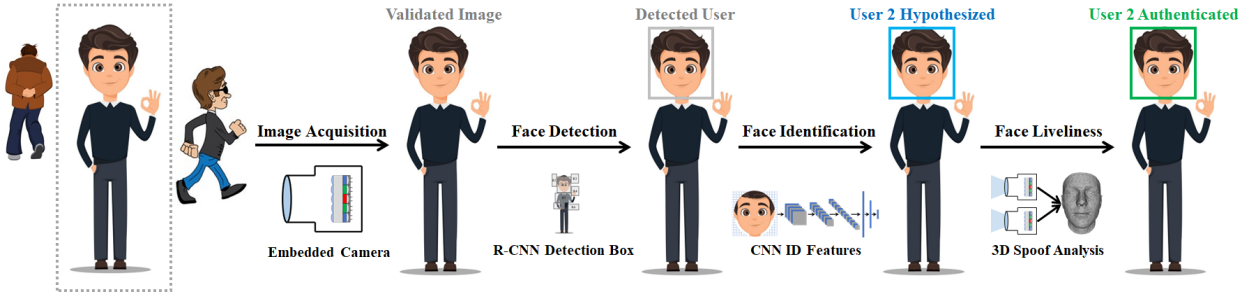


Figure 2.1: Traditional face-recognition pipeline. Most platforms acquire an image, then run detection, identification and liveliness algorithms in series.

2.1.1 Image-Acquisition

Image acquisition is the process of acquiring a valid image from the camera and correcting any relevant distortions. Real-time-systemss typically use an embedded-camera that has a direct interface to the processing module. While possible to do off-board reasoning distribution (e.g., cloud processing), this is usually not done to avoid exposing biometric data.

The acquisition process usually has some basic signal processing to yield a quality image. This usually includes automatically adjusting the exposure and gain based off lighting levels (28), as well as demosaicing and interpolating pixel channels (29). If a fish-eye lens is utilized, relevant distortion correction is often applied (30). These processes can be done with either the reasoning module or an image-signal-processor within the camera. Encryption is only recently being applied due to authenticity requirements (31; 1). Many existing systems only do basic source-replay mitigation to save on compute (exposing the photo-realistic tampering vulnerabilities.)

2.1.2 Face-Detection

Valid images are analyzed for faces. The purpose is to precisely localize faces and segment them from the background (de-noising identification and liveliness features). Historical methods employ gradients analysis, such as Viola-Jones (32) and histogram-of-oriented-gradients (HOG) (33). Current state-of-the-art is to employ deep-learning (DL), inferring detection anchor boxes from regional proposals. Region-based convolutional-neural-networks (R-CNNs) (34) and feature-pyramid-networks (FPNs) (12) are particularly popular today. R-CNNs offer a means to efficiently evaluate detection location hypothesis and FPNs incorporate information from multiple feature dimensions to improve localization accuracy.

2.1.3 Face-Identification

Once a face is detected, it is determined whether it belongs to an enrolled user. Facial-identification methods have historically started with matrix (35) and texture descriptors (36). Current state-of-the-art methods now implicitly describe the features using DL. The challenge here is be able to differentiate one person from over 50,000 per industry standards (1). To achieve this precision, algorithms often employ sophisticated loss functions. Popular examples include FaceNet via triplet loss (37) and ArcFace via cosine loss (13). These DL algorithms are utilized to describe incoming faces with embeddings, which are then compared against the enrolled users for similarity. If the similarity-score for a given user passes the evaluation threshold, it is presumed they are the person in the image. Note that similarity-score threshold determination is an imperfect science, usually done experimentally on the test-dataset to get the desired accuracy metrics.

2.1.4 Face-Liveliness-Verification

Lastly, the identified face is verified for liveliness. Imposters can produce spoofs (facsimiles) of enrolled users to bypass the FR security. Depth is the preferred approach at this time (18). With that said, others are investigating if materials can be identified from texture (38) and their temporal (15; 16; 39) behaviors. This section is greatly expanded upon in Chapter-4.

2.2 Application Services

FR platforms typically provide a few fundamental services. In addition to authenticating people, there are profile management services. These often consist of adding accounts, managing the levels of authorization and removing accounts.

2.2.1 Face-Enrollment

Enrollment service adds a new valid face-profile to the user database. Users typically need to first authenticate themselves (using another technology) to initiate the enrollment. From there, they are often guided by a human-machine-interface to perform a series of poses for profile creation. The enrollment robustness (e.g., perspective variance) usually improves authentication reliability, but can feel tedious if too many poses are required.

2.2.2 Face-Authentication

Authentication service verifies whether a presented face belongs to an enrolled user. This follows the pipeline presented in Fig. 2.1. A person is considered authenticated if their face meets identification similarity and liveliness requirements.

2.2.3 Face-Authorization

Authorization service verifies whether the authenticated person has sufficient permissions associated with their profile. Note that authorization is typically done on the application layer, but is directly associated with the authenticated face.

2.2.4 Face-Removal

Removal service removes a valid face-profile from the user database. This means the face-profile can no longer be used for future authentication. Typically, users can only remove their own profile; removing other people's profiles requires a high level of authorization.

CHAPTER 3

Face-Recognition Threat-Model

This purpose of this threat-model is to elucidate face-recognition (FR) vulnerabilities. Attackers can leverage the sensor, algorithms and even the user as illustrated in Fig. 3.1. The first use-case is the intended one, an enrolled user looking to authenticate - no threat analysis required. The remaining are attacks: replay, spoofing, face-swap, denial and phishing. Note how the general trend is to gain unauthorized access, but one can also deny access (or generally control the application with phishing credentials). This chapter evaluates these threats with risk assessment.

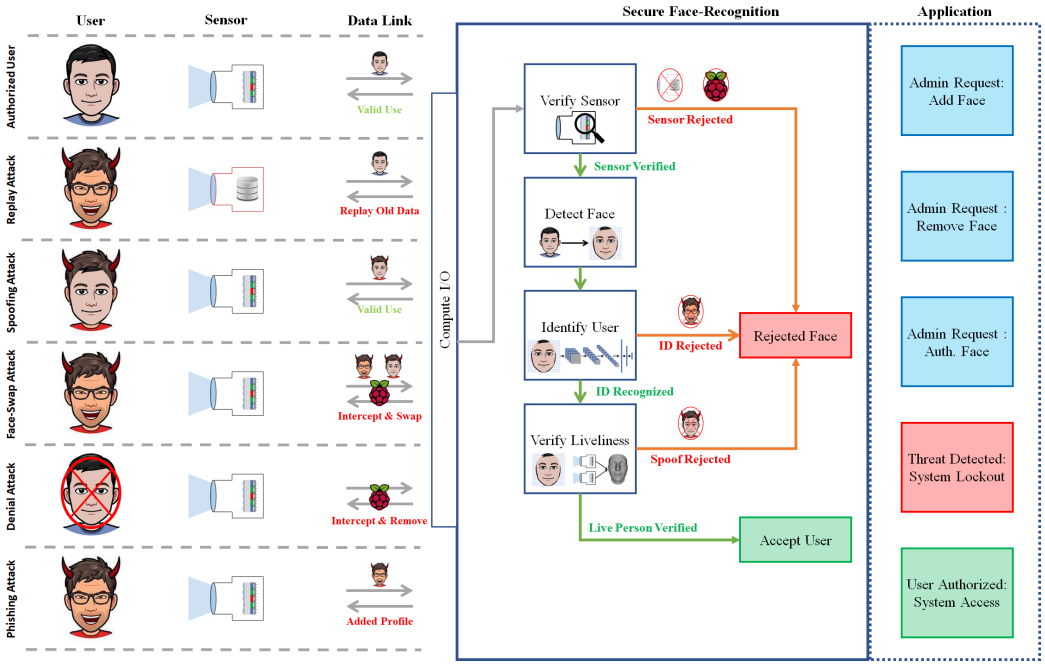


Figure 3.1: Face-recognition threat-model. Security counter-measures are visualized.

3.1 Replay-Attack

Replay-attack is the most fundamental method of gaining unauthorized access (threat-model use-case 2). The attacker records a valid data-stream from the camera, then replays it at a later time for authentication (40). This can be mitigated, however, by using basic networking protocols. Common practices include message authentication (41) and watermarking (42). Given this problem is largely addressed, there is no further consideration given.

3.2 Physical-Spoof-Attack

The physical-spoof-attack (PSA) is essentially a facsimile based "replay" attack (threat-model use-case 3). Here, a physical representation of an enrolled user is presented in lieu of replaying sensor data. For reference, the National Institute of Standards and Technology (NIST) has an international standard for facial spoof presentation. ISO 30107 has three proposed levels for PSA: Level A, pictures; Level B, video replay and paper masks; Level C, 3D masks (14). This chapter builds off the NIST standard and separates Level C into two categories based off cost and impact on detection. Spoof vulnerability is derived from success rate, juxtaposed with production time and cost. The goal is not to be impervious, but rather dissuade attackers by sufficiently detecting all easily-produced vectors. Common spoof presentations are visualized in Fig. 3.2.

Industry guidelines suggest secure applications should target a 5% attack-presentation-classification-error-rate (APCER) (14; 1) and convenience applications 20% APCER (1). These ratings are expected-value and should reflect presentation frequency. Given attackers traditionally go after repeatable, lower-effort methods, algorithm developers are highly encouraged to achieve stringent APCER values for 2D spoofs. 3D spoof detection is usually not a concern except when the risk of failure is particularly significant (e.g., government facility access). Note while the verification of live people is not a security problem, having a poor rejection-rate will result in user-dissatisfaction.

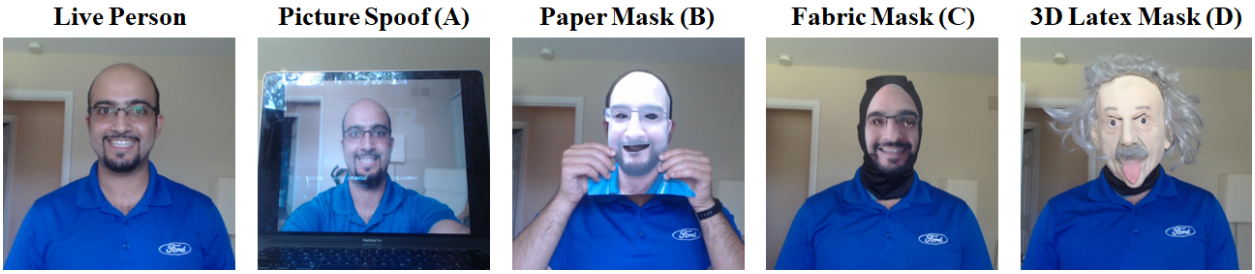


Figure 3.2: Facial-spoof presentation illustration. Common attack presentations include still picture, 2D mask or video, simple 3D mask and highly realistic 3D mask.

3.2.1 Level A: Still-Picture

A still-picture spoof is the presentation of still picture of an enrolled user (14). This picture can be easily acquired by access to the user's social media, employment profile or personally taking a picture of the user (14). Given the benefits of having a quality head shot photo online, time to acquire a quality image is often negligible. The picture may be either printed or on a display (such on a smart phone), such that cost is also typically negligible. It is by far the simplest attack vector, and is often the first to be conducted by the media (see Samsung S10's spoofing exposure as an example (43)). Level A spoofs are almost certain to be presented and must be mitigated to ensure user trust.

3.2.2 Level B: Video-Replay

A video-replay spoof is the presentation of a video containing an enrolled user (14). The video can similarly be acquired via access to the user's social media platform or personally recording it, though there is inherently a bit higher barrier to entry due to potentially needing post processing (14). That is to say head shots are relatively easy to acquire, but a well-focused video that is appropriately zoomed in on the user is not. It is likely the attacker would instead have to gain access to a video that includes a desired clip, extract the target frame, and then display it on repeat on a smart device. Level B spoofs require a bit more skill and effort, but are also very likely to be presented. These also must be mitigated to ensure user trust.

3.2.3 Level C: Simple-Mask

A simple-mask spoof is the presentation of a mask representation of an enrolled user produced from 2D imaging, where the face is cropped, and the eye and mouth holes are cut out (14). Common mask materials are paper and fabric, which can be produced for under $20; note that latex is also an option, but is typically not used for cost reasons (14). Production effort depends on the material. A paper mask can be created with an hour of effort; this consists of printing the image and then trimming it to remove the outer boundary and expose the eye and mouth holes. Fabric masks require more slightly more effort, as now the user must iron on the print out onto the fabric. However, there are also online vendors who create fabric masks for decorative purposes (44). Online services trivialize create - simple upload a photo - but does introduce the time associated with commercial production and shipping.

This spoof is specifically designed to affordably defeat state-of-the-art systems. It involves incrementally more production time, but has the capability to defeat many systems that introduce depth and motion. Level C spoofs are considered the greatest threat to FR.

3.2.4 Level D: Highly-Realistic 3D Mask

A highly-realistic 3D mask spoof is the presentation of a detailed mask representation of an enrolled user produced from 3D image, where again the face is cropped with eye and mouth holes cut out (14). Barrier to entry is considered extremely high. Acquiring a 3D model of the enrolled user can theoretically be done with an eclectic assortment of images from various angles, but more realistically should be done in a controlled recording studio, with the assistance of a software expert to generate the model (14). Custom 3D masks production can easily cost thousands of dollars and usually takes months to fabricate. When considering these factors, the highly realistic mask is not likely to be reproduced by except by those who have significant time, resources and motivation to defeat a given system.

Detection Level D spoofs can be extremely challenging. A 3D generated mask can fool almost all current in production system, however, due to production complexity are rarely presented beyond academic research. Level D spoofs are out of scope for this reason.

3.3 Face-Swap-Attack

The face-swap-attack (FSA) is a modification of valid frames (threat-model use-case 4). By digitally swapping in a face of an enrolled user and using photo-realistic-blending, FSA can easily fool state-of-the-art identification algorithms (26).

This vulnerability is made accessible to insufficient network security. For efficiency purposes, most embedded systems do not encrypt image streams; instead, message are authenticated using a combination of digital signatures and meta data (41). Furthermore, many watermarking schemes are designed so as not to manipulate the principal data (42); hence, the security systems are often tailored to watermark pixels that do not contain the face. Current detection methods are computationally expensive (typically re-current neural-networks (45)) and not a good fit for FR systems. The FSA is a key vulnerability for these reasons.

3.4 Service-Denial-Attack

The service-denial-attack (SDA) prevents valid users from authentication (threat-model use-case 5). The primary utility is often to cause inconvenience. By interfering with the FR system, the user is forced to use an alternative authentication method. Secondary utility can include blackmail, whereby the attacker demands payment in exchange for resuming service (46).

The simplest method is to intentionally trigger the system lockout. It is common for FR applications to implement a time delay between repeat failed-authentications, and ultimately prohibit future attempts until authorized by another factor (47). By repeated presenting invalid faces, the attacker can intentionally fail until they are prohibited from additional attempts, effectively locking out valid users. Note this attack is common to many authentication modalities and not unique to FR. For this reason, it is viewed out of the liveliness scope.

A more sophisticated method is to digitally remove the face from the images in real-time. Similar to the FSA, an attacker can place an interception device and photo-realistically remove the faces to prevent authentication. This attack can also be used to falsely trigger alarm systems by swapping in faces of blacklisted people. Detecting face-removal has the same issues as face-swap (complicated deep-networks (45)). The SDA is also a key vulnerability for these reasons.

3.5 Phishing-Attack

The phishing-attack is designed to gain credentials by deceiving enrolled users (threat-model use-case 6). The attacker typically acquires credentials through deceptive messaging that encourages people to verify their account (providing their credentials in the process). This primary intention is generally for the attacker to function as system administrators. This provides the capability to either enroll themselves or deny access by removing others.

Phishing-attacks are not viewed as a liveliness issue. There is no replay of valid users present and can solved by applying best-practices with the system administrators. For example, simply adding two-factor authentication for profile management secures these attacks. They are considered out of scope for these reasons.

3.6 Addressing Facial-Liveliness

This threat-model shows there are numerous ways to manipulate FR applications. This research proposes the ones most relevant to liveliness are the PSA, FSA and SDA. More pertinently, the PSA is the most likely one to present in real-world attacks. Image-tampering attacks necessarily require physical access to the sensor feed and access to the data. This is notably more complex than simply presenting a picture or video to the FR system. The majority of this dissertation is aimed at mitigating PSAs for these reasons, where one-chapter addresses both FSAs and SDAs.

3.6.1 General Security Practices

In addition to developing a robust facial-liveliness-verification methodology, it is beneficial to also follow general security best practices. One well represented practice is to assume the system is vulnerable, and require a periodic secondary authentication. This is typically done on smart devices where the user must verify a password either after so many biometric authentications, or days elapsed since their last confirmed password. Another is to enforce a progressive failure back-off. Brute force attacks can be limited by requiring an increasing timer in between failures; a permanent lockout can also be introduced after so many consecutive attempts. Attackers can be dissuaded by simply slowing down the rate at which they can identify vulnerabilities.

CHAPTER 4

Facial-Liveliness-Verification State-of-the-Art

Face-recognition (FR) is a rapidly advancing field. This technology started in the 1970s with simple matrix representations (35), steadily growing into the deep-learning pioneer it is today (48). Significant advances have been made particularly in the detection and identification spaces. Detection algorithms can now localize the tiniest faces, even when rotated or occluded, by using feature-pyramid-networks with facial-landmarks multi-task-learning (12). Similarly, identification algorithms can now achieve almost perfect accuracy on competition celebrity-datasets by using introducing sophisticated contrastive-losses (37; 13). The area that still requires fundamental methods exploration is monocular (efficient) facial-liveliness-verification (FLV).

The threat-model helps elucidate that the primary FR vulnerabilities are the spoofing and image-tampering attacks. The primary goal is often to "replay" an enrolled user, either in the form of a facsimile or digital manipulation, to gain unauthorized access. Secondary objectives may include denying service, where the attacker digitally removes faces from the camera stream. For more details on these attack methods and other vulnerabilities, see Chapter-3.

This chapter presents the corresponding FLV state-of-the-art. Relevant works are presented and then analyzed with their capability to meet the monocular, single-frame objectives. Survey is performed using Google Scholar, IEEE Xplore, IEEE Transactions on Information Forensics and Security, industry requirements and international standards as references. A breakdown of these works is given in Table 4.1.

Method	**Papers Identified**	**Papers Cited**
Physical-spoof-attack detection	200	40
Image-integrity-verification	100	39

Table 4.1: State-of-the-art survey metrics. All other papers cited in this paper are either relevant face-recognition background or tools used for development.

4.1 Physical-Spoof-Attack Detection

The physical-spoof-attack (PSA) is a facsimile attack typically derived from 2D imagery (14). These types of attacks, such as display a video or creating a mask from a paper-printout, can be done with a trivial amount of production effort and cost. The risks here are ever more present as people take more advantage of social media and share quality imagery online.

Ultimately, the goal is to identify methods that can deployed on mass. In principle, ultra-precise 3D sensing is the ultimate standard with facial-liveliness-verification (FLV) (49). This research does not debate this, but rather notes the intended use-case (highly-realistic 3D masks) is often unrealistic. In addition to being expensive (surveyed local costume companies charge between $5,000 and $10,000 USD), it requires a precise 3D model of the intended victim. While potentially feasible from 3D reconstruction methods, it is ultimately viewed as very unlikely. Once factoring in the cost and computational needs of such 3D sensing systems, it is more pragmatic to focus on mass-usability methods.

Hence, emphasis is placed methods that mitigate 2D inspired spoofs and are capable of mass-deployment. This survey identifies the relevant state-of-the-art, noting there is a literature gap with respect to monocular, single-frame algorithms. A discussion is presented on how the proposed material-spectroscopy approach can address this.

4.1.1 Depth

Depth is introduced first as it is generally viewed as the best approach to mitigating 2D inspired spoofs (49). This is rather intuitive, as short of having a 3D model of the intended target, there will always be some artifacts when making a facsimile. Even projecting an image onto a 3D mannequin will necessarily change the 3D structures to align with the mannequin. Hence it is important to discuss depth methods even though they fail the monocular requirement, as they give an indication of how to best approach PSA detection.

Depth maps are historically imaged using passive stereo vision, where two cameras are placed in a known relationship, and depth is calculated by examining the disparity between the images (50). This offers an avenue into detecting simple masks – though is fundamentally limited. Passive stereo vision typically isn't very precise without very high-resolution cameras and often has a limited operational range; for these, it is preferred to use structured light. Structured light systems project a known pattern, often in the form dots or dashes (51). This light pattern can then be used to triangulate points on the face to calculate a very precise depth map - aka active stereo; if the light pattern is serialized, it can also be done with a single camera (aka Apple's Face ID). This is not computationally cheap, but can result in precise depth maps.

If cost is not a concern, structured light has a lot of promise. It can be precise enough to reliably detect flat spoofs (pictures and replayed videos) (18), with promise to detect simple masks without much curvature. One major current constraint, however, is the field of vision tends to be very narrow. Current outdoor grade emitters for security and automotive often cap at 20-degree field of vision. In principle this should be improvable with better emitter design - but in practice it means current systems require sophisticated feedback mechanisms to help the user align with the camera system.

4.1.2 Motion

Motion methods are rather popular for addressing the PSA when depth technology is not available. Many spoofs are inherently rigid; whether a picture or a latex mask, the facial structures are constrained in place. This rigidity can be quantified using a variety of temporal algorithms. Hence it is also important to discuss motion features, despite that they fail the single-frame requirement. These can serve as a relevant benchmark as to what expectations can be placed on monocular systems.

Eye tracking is the simplest way to detect a still picture spoof. A simple blinking by tracking the landmarks around the eyes can be used to counter these attacks (15). This is computationally cheap way to mitigate picture attacks without any impact to cost or user experience. That said, eye tracking should be always used in concert with more sophisticated methods, as it is trivial to defeat. A video replay inherently provides blinking, and an attacker can simply cut eye and mouth holes in a picture to make it a paper mask.

In the same vein of motion analysis, there is state-of-the-art research into how the various muscles of face naturally behave. For example, the mouth must perform a predictable series of contortions to form speech; even breathing predictably contracts and relaxes the cheeks. This can be detected by training temporal deep-learning (DL) networks (52), where isolating the behavior of the mouth can improve performance (53).

This adds an extra layer of security over eye blink detection, as now the attacker can no longer simply cut eye holes, but must fashion a mask that closely follows their facial movements. Performance can be notably increased if a microphone is also included, where the facial motions can be synchronized with audio, where an expected behavior can be predicted for comparison (54). Still, natural micro-motion is best served as a component in an ensemble of countermeasures. It is similarly vulnerable to video replay, and there is particular risk of requiring motion as well. A system that is calibrated sufficiently to reject motion transmitted from simple masks will likely reject still people. Incorporating cooperative gestures would certainly help - but that would beg the question whether this step is necessary.

Heart-rate detection is another method that can be included in an ensemble anti-spoofing method. Blood flow across the face is periodically pulsing with heartbeat (16); each time the heart contracts, the veins across the face will similarly contract. This is detectable using frequency domain analysis to track the contractions (16). It can be also detected by a shift in the green channel as the blood's hemoglobin absorbs green light (55).

Heart-rate tracking in principle shares a lot of the benefits of natural micro-motions. It is very difficult for even a highly realistic mask to produce a heartbeat, and a video will similarly pass these detection methods (56). One could argue that an attacker simply needs to cut out a forehead patch of the mask to expose their own heartbeat. Regardless, motion sensitivity is a fundamental challenge. This paper's evaluation of both methods takes multiple seconds to acquire a stable signal; if the subject is moving, it is often impossible. This is reflected in the original works by referencing an average over a 30-second time period in the results (55). Heart is a quality feature when available, but the sensitivity to motion suggests it is similarly best used in an ensemble.

4.1.3 Texture and Color Artifacts

When constraining the methodology to monocular and single-frame, PSA can be detected through reproduction artifacts. For example, printers can predictably add distortion and quantize the colors of the intended face (38). Likewise, both material and geometry can bias the way the spoof interacts with light; this can bias the distribution and therefore the perceived texture (38). From a theory perspective, these types of artifacts can be identified from image texture and color accuracy.

One of the popular facial-texture descriptors are local-binary-patterns (LBP) (57). The LBP describe the local relative rate of change, or gradient, for a given image patch. This is particularly useful to generally differentiate between faces; however, Chingovska et al. rather famously demonstrated LBP cannot differentiate liveliness on color imagery (58). This is believed to be a shortcoming of using a passive RGB camera instead of an illuminated near-infrared camera.

While not common to facial anti-spoofing, a texture descriptor used in fingerprint anti-spoofing is the Gabor filter (59). Gabor filters describe images in a sinusoidal fashion, incorporating a wavelength and phase into the kernel (59). It is commonly used for edge detection applications and postulated that it could similarly be used for facial analysis. Note that these methods are demonstrated only on paper and display spoofs. It is still an open area of research as to whether more sophisticated spoofs could be detected.

Color distributions can also theoretically identify spoof production artifacts (60). Wen et al. have demonstrated that ranked-channel-histograms can in fact capture the color artifacts, but their results are only shown to work under static lighting. When reproducing their work, this research finds this approach to be too sensitive ambient lighting.

4.1.4 Texture and Motion Fusion

Given the shortcomings of artifact detection for the PSA, recent methods combine texture and motion. In theory both provide relevant features that are individually insufficient but their fusion can serve the goal. This can be done very explicitly, such as take a sequence of texture maps (61). Others have taken a slightly more elegant approach and use spatio-temporal networks, using temporal layers on a texture network to infer features (62; 63). These methods are generally accurate, but are computationally-complex and do not meet the single-frame requirement. Furthermore, it is an open question whether temporal networks are robust against natural motions. Most evaluation datasets keep the spoof mask relatively still; quickly shifting the mask may be problematic.

4.1.5 Non-Visible Spectra (CMOS)

Attackers see in the visible spectrum, and therefore design facsimiles to be identical to their own perspective. This can be potentially advantageous, as a material reflectivity is not necessarily consistent across different light spectra. By changing the camera spectrum to go beyond visible light, one can introduce facsimile artifacts.

Near-infrared (NIR) is a great spectrum to detect PSAs. It introduces variations in reflectivity and is relatively eye-safe in comparison to ultra-violet (the other option). Many common spoofing materials appear brighter to NIR cameras. Furthermore, most displays only emit in the visible spectra and do not show up. These phenomena can be seen in Fig. 4.1.

Figure 4.1: Juxtaposing human and illuminated near-infrared perspectives. Spoofs have less texture variance and appear brighter to the infrared camera.

In theory, this filtering works on any non-visible spectra. CMOS is suggested for cost reasons. It is the most commercially available photo-receptor and hence has the high resolution necessary for precise facial identification. Common photon quantitative emission curves show that CMOS can both perceive some ultra-violet and some NIR. Either spectra can offer unique insight into the facial liveliness. Any illuminated systems, however, should use NIR for eye safety purposes.

Skin-reflectance is a way to effectively characterize the "skin" material. This is a form of material-spectroscopy (MS), where a controlled light is emitted, and the material is identified based upon the reflectance pattern. This is a known material property, where certain spoof materials can be identified based off their reflectance coefficient (64).

The challenge with spectroscopy is controlling for the environment. The conditions represented in Fig. 4.1 are ideal. There is minimal ambient light, the distance is controlled, and it is a direct comparison of a known person and their paper mask (same skin tone). If these factors can be controlled, spectroscopy has the benefit of being implementable with CMOS cameras and standard flood illumination.

4.1.6 Non-Visible Spectra (Non-CMOS)

Temperature is arguably the most robust method to detect any PSA. A live face will have natural gradients in temperature as a result of the blood vessels underneath (16). Skin patches are directly above blood vessels will be warmer, with a pulsatile behavior that is also directly a result of the heart's contractions (16). This thermal-pattern is extremely difficult to reproduce on spoof materials. Pictures, videos, simple masks and highly-realistic latex masks can be detected with thermal analysis (65; 66). This liveliness-contrast is visualized in Fig. 4.2.

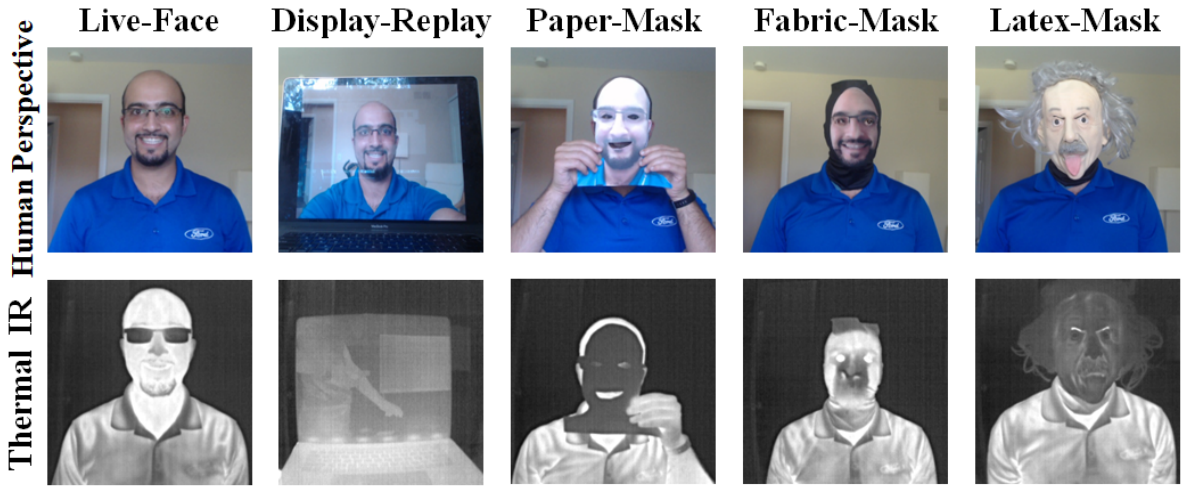


Figure 4.2: Thermal liveliness visualization.

Long-wave-infrared (LWIR) (thermal spectrum) cameras work off radiated light. Note how even though some heat is transferred from breathing, it still looks markedly different. This technology is generally power efficient, as there is no need to illuminate. However, the lack of illumination means thermal imaging can lack contrast.

A potential compromise between NIR and LWIR is to go in between the wave bands via short-wave-infrared (SWIR). SWIR is a longer, reflected infrared waveband. This offers an opportunity enhance upon the reflectance analysis of NIR by picking a wavelength that well differentiates materials while also having imaging contrast (67). Due to cost, very little SWIR is currently published, but this can be potentially a very effective perception technology.

4.1.7 Cooperation

User cooperation can be another implicit way to mitigate PSAs. Some level of cooperation should always be required regardless, as a way to establish intent. Because facial recognition is by design a passive authentication method, there are both security and consumer trust benefits to only act when the user has expressed an intention to use said system. A basic example would be requiring mobile phone access to have eye gaze; one does not want to risk their children unlocking their phone when they sleep, and frankly there is a bit of a creepiness factor if the phone automatically unlocked every time they were simply nearby. As such, it is very common to require either a cooperative head pose or gaze before taking any control action.

Introducing a challenge gesture can be a great way to incorporate a second factor while retaining a clean user experience. The user can be prompted to smile, frown, laugh - any gesture is that difficult for a mask to reproduce will suffice. Even when considering highly realistic masks, it is challenging to produce a mask thin enough that it can contour with the attacker's facial muscles well. The primary downside is this requires sophisticated feedback. One cannot reasonably expect the user to perform an arbitrary gesture if they are not shown prior. Audio prompts are also recommended (it is challenging to both read a display and look at a camera at the same time).

4.2 Image-Integrity-Verification

Image integrity-verification is a fundamental scope of image forensics. In his famous Image Forgery Detection survey, Dr. Hany Farid presents how images are commonly tampered with corresponding detection methods (19). Images can be tampered for various reasons. People have been known to present fake imagery to forge alibis (68), steal identities (69) or create compromising material (70). This can be done using a variety of handmade tools, such as: cloning, duplicating parts of the image to conceal or embed information; re-sampling, adapting the resolution to mod-

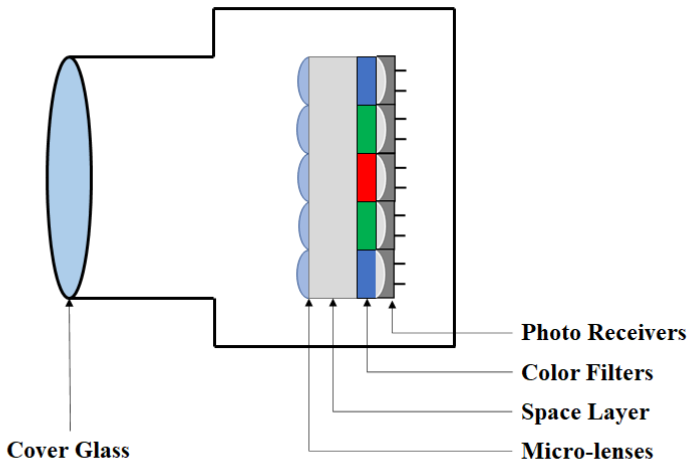


Figure 4.3: Illustrating camera components.

ify scene proportions; and splicing, combining multiple images together to create a new scene (19). Today these effects can be done photo-realistically using DL networks (71). State-of-the-art face-swap methods also enable noise blending for highly realistic forgeries (71; 70); generative-adversarial-networks (GANs) are in particular popular as they are easy to implement and designed to look photo-realistic (72).

4.2.1 "Noiseprint" Based Integrity-Verification

To detect these image attacks, many methods typically rely on noise artifacts for image forensics. The camera's fundamental components are illustrated in Fig. 4.3. Note the lens (or micro-lenses) and photo receivers in particular. Manufacturing tolerances make it nearly impossible to create a perfectly noiseless component. Every lens will inherently focus light slightly differently, as every photo-receiver will receive light slightly differently. When aggregated over all the pixels in the camera, this can translate into a robust "noise fingerprint" or "noiseprint" (19).

The "noiseprint" of choice here is photo response non-uniformity (PRNU, sometimes referred to as fixed pattern noise). This is an estimation of photo receiver noise with respect to a constant uniform light (19). PRNU is used in particular for camera source identification because of its relative independence from ambient lighting. Lens aberration, the offset in focus due to lens imperfections, is difficult to quantify because it requires information on the light source (19). PRNU, on the other hand, has been shown to be robust by various research groups over millions of images (73). This includes encoding and compression introduced from social-media platforms, the default means of acquiring quality tampering imagery (74).

Dr. Jessica Fridrich's group (75) has conducted landmark research for using PRNU in camera source identification over numerous conditions. Their methodology first extracted the PRNU by removing temporal noise and estimating the noise residuals and then determining camera source via peak-correlation-energy (PCE) (75). In particular, they demonstrate robustness scales over millions of images from thousands of cameras, including tolerance to the compression used by the online image repository that they data mined (Flickr) (73). They note that feature margin is decreased when using images of the same scene, in particular from the same camera model, but in general PRNU is reliable for source identification (73).

Based on these findings, it is natural to hypothesize that PRNU can be leveraged to detect image tampering. Others have begun exploring this hypothesis. Chierchia et al (76) demonstrate that PRNU can be used to detect when a picture clipping has been injected into the image plane, so long as the clippings come from different scenes taken by different cameras. Korus and Huang greatly expand on these results by using a multi-factor approach, combining multiple candidate tampering maps and applying stochastic segmentation to dynamically determine the analysis window dynamically (77). This method brings noted improvements in reliability, demonstrating the capability to detect tampering down to local square regions (77). Their tampering window analysis indicates that 128 pixels square is mostly robust, where 256 pixels square can achieve desired targets of at least 90% true positive rate for 5% false positive rate (77), though requiring semantic segmentation (computationally heavy).

Note that while PRNU is one type of "noiseprint", other noise factors can be used. Mahdian and Saic use discrete wavelet transform analysis (78). Levandoski and Lobo similarly demonstrate an ensemble of duplication, color filter array and noise detectors in their Forgery Tool (79). While these are not PRNU detectors, they operate off the same premise and have available source code; for these reasons they are used for competitive benchmarking.

4.2.2 Deep-Learning Based Integrity-Verification

This paper emphasizes "noiseprint" analysis due to the emphasis on real-time activity. With that said, when it comes to detecting photorealistic imagery, DL methods represent state-of-the-art performance. In some cases, noise features are pre-processed. Cozzolino and Verdoliva use a transformed PRNU for pixel-level tampering localization (96-97% accuracy) (80). They then extend their own by training a Siamese convolutional neural network on native and tampered image PRNU for improved performance (81). Gunawan et al similarly use error levels analysis (ELA) (82); their work is used as a benchmark as they have provided source code. It is relevant to note, however, that the images for these papers are tampered using traditional splicing methods.

Alternatively, networks can be trained on residual traces from the generator. Guarnera et al trained a network on images generated by numerous GANs, implicitly inferring noise artifacts (83). Their performance is very strong, ranging between 88% and 99% (depending on the GAN). The one major constraint in this approach is requiring the detector to be trained on the specific fake image generator. If a new GAN is developed without the detector's knowledge, they will have reduced security.

A recent theme is to use context over a multitude of frames. While it is possible to deepfake a single frame, it becomes more challenging to fake natural movements. This is then detectable using recurrent neural networks (e.g., incorporation of long-short term memory over a sequence of frames). Guera et al demonstrated very strong performance, hitting 96 % accuracy on test videos when using 20 frames (45). Others have similarly pursued multi-frame analysis. Sabir et al demonstrate that strong performance can be achieved with fewer frames if aligning the faces (84); however, Tariq et al note large scale robustness over varied perspectives does benefit from increasing the frame count (in this case 16) (85). This approach seems promising, but is inherently not ideal for real-time applications due to acquiring numerous frames.

4.2.3 Application to Face-Swap-Attack

With all this said, one of the key questions is whether the face in the image can be trusted for identification. The face-swap-attack (FSA) can defeat even the best models (26), where even the reliable FaceNet (37) has been successfully fooled by digitally swapping in the face of a properly enrolled user (26). Today, face-swapping algorithms are used primarily for entertainment purposes, with many social media applications allowing users to do swaps for fun. This, however, also provides financial incentive to develop better swapping algorithms. Speed (86) and image quality (87) are both industry metrics; the unintended consequence is implicitly developing toolboxes for attackers to robustly and maliciously tamper facial images.

Given that facial manipulation is one of the first applications for GANs, the state-of-the-art in FSA detection overlaps with deepFake detection. For example, all of the recurrent neural network applications cited previous included face-swap data sets in their evaluation (45; 84; 85). Another relevant example is the work of Rossler et al, who provide a public data set of tampered face videos, FaceForensics++, along with their own detection network (88). Their network, Xception-Net, performs very well (99% labelling correctness, including 81% on compressed data) (88) and has inspired numerous other high performing models, such as using multi-task learning (89) and optical flow (90).

This shows that there is promising performance in the domain of video. One can intuitively understand that faking photo-realism is an easier challenge than faking realistic human movements.

As such, there is work to do in the domain of single-frame analysis. Recall again the objective is to mitigate photo-realistic tampering in an imperceivable fashion.

4.3 New Methods Exploration

There are a number of related works that address spoofing and image-tampering attacks. The challenge at hand is meeting the monocular and single-frame goals. Recall these goals are selected to make the methodology mass-deployable. While these constraints effectively eliminate all of the identified methods, they can still be used as sources of inspiration.

4.3.1 Physical-Spoof-Attack Detection

The general theme to detection the PSA is identifying facsimile artifacts. Given depth features are clearly ideal, the obvious question becomes how to acquire them in a monocular, single-frame fashion? One hypothesis is to utilize reflected light. Object geometry will dictate the reflectance patterns meaning some depth information should be available. This information, however, is inherently more noisy than using a dedicated sensor. To improve signal, it is observed that changing the light spectrum to near-infrared can introduce material separability (as attacks are designed to look realistic to the human eye).

Hence, the proposal is to utilize reflected near-infrared light to characterize the material (e.g., material-spectroscopy). There are observable differences between live and spoof faces in the near-infrared spectrum as visualized in Fig. 4.1. It is intuitive that this approach can work for 2D inspired spoofs so long as the right texture descriptor is identified. Should highly-realistic 3D mask detection be necessary, it is proposed to employ thermal imaging. The contrast between live and spoof faces becomes even more stark in this spectra as visualized in Fig. 4.2. This should easily translate into a monocular, single-frame algorithm. The only issue is cost, noting thermal cameras are potentially powerful but expensive. As such, the dissertation will focus on near-infrared spectroscopy.

4.3.2 Image-Integrity-Verification

General image tampering detection can be done using camera noise analysis. Noise profiles are difficult to fake and will be necessarily modified by any type of tampering attack. Dr. Luisa Verdoliva's group has in particular advanced this field and demonstrated various "noiseprints" that can even be used on photo-realistic forgeries (81). It is proposed that this type of noise analysis can be applied to verify detected faces are authentic (mitigating the FSA) as well as the full image is authentic (mitigating the SDA).

The challenge at hand is the two-fold. First, the advent of generative-adversarial-networks means that new forgery methods are being released at a rapid pace. Even if some generators artifacts are well understood, it does not necessarily guarantee robustness to future ones. Furthermore, the methods identified are all using sophisticated deep-learning (computationally-expensive). Hence, the proposal here is to leverage the fact face-recognition systems typically own the imaging system. The camera's noise profile can be characterized, simplifying the problem to simply detecting anomalies against the enrollment. This is believed to be sufficiently accurate and computationally efficient.

CHAPTER 5

Proposed Frameworks

The goal at hand is to achieve robust facial-liveliness-verification (FLV) in a monocular, single-frame fashion. These goals are designed for mass-deployment across face-recognition (FR) systems. The primary vulnerability is the physical-spoof-attack (PSA). It is rather trivial to find a headshot of the intended victim online, then present a facsimile for authentication. Secondary vulnerabilities are DeepFake based, and include the face-swap-attack (FSA) or service-denial-attack (SDA). These can photo-realistically present the intended victim or altogether remove faces from the scene. Note, however, these require interception device placed at the sensor and are inherently more complicated. For more details on vulnerabilities and attack methodologies, see the threat-model in Chapter-3.

This chapter introduces a series of physics-informed frameworks to address these attacks. For example, the PSA is typically performed using facsimiles inspired from 2D imagery. This should necessarily introduce differences in geometry (and potentially material-reflectivity). Likewise, any sort of image tampering (e.g., FSA and SDA) should necessarily alter the sensor's intrinsic noise profiles. This intuition is applied towards identifying monocular features that are computationally efficient. This is later validated with mathematical modelling and experimentation in the associated framework chapters.

5.1 Anti-Spoofing via Near-Infrared Material-Spectroscopy

Spoofs are artificial-facsimiles inspired from 2D photos (14). Common presentation methods include pictures, videos and simple masks (noting that 3D masks are possible but unrealistic due to requiring a 3D facial-model) (14). These presentations all lack the 3D robustness of a live face. As such, depth features are typically best at detecting them (49). Depth features can be acquired from 3D sensors (91) or multi-frame deep-learning networks (92). The goal is not necessarily to supplant those methods, but rather demonstrated similar robustness can be achieved without the cost and computational overhead.

Figure 5.1: Visualizing common spoofs in RGB versus NIR. Presentations in order: live-person, paper-mask and display-replay.

Material-spectroscopy (MS) is the process of identifying objects by shining a controlled light source and characterizing the reflection (93). Humans see in visible light, and therefore design spoofs to look identical to their own perspective. This similarity, however, does not necessarily translate across different light spectra. For example, a paper-mask that looks realistic in the visible band appears bright and washed out in the near-infrared (NIR) band. This reflectance phenomenon is visualized in Fig. 5.1.

This research proposes in Chapter-6 that texture methods can robustly characterize liveliness from NIR reflectance-patterns. The observed differences are a function of the object's geometry and material; hence, it is proposed that the geometry can be inferred from reflectance texture. This is formalized with a mathematical model.

Live faces are a combination of multiple surfaces with varying radii of curvature. The given mathematical model shows this must necessarily generate a distribution of varying-frequencies (as a function of radius of curvature). Conversely, the spoofs considered are either flat or essentially a simple convex-surface. The given model also shows how this results in a simple low-frequency distribution. These distributions are further biased by the material, where spoofing materials are essentially uniform and very reflective (e.g., paper, fabric and displays) and live faces are composed of skin, hair, eyes, etc. that vary in reflectivity. The proposal is validated using a large-scale experiment, demonstrating a panel of texture classifiers (employing both deterministic and deep-learning methods) can robustly verify liveliness.

5.2 Addressing Spectroscopy Sensitivity to Camera and Environment Noises

The material-spectroscopy algorithms infer geometry and material-reflectivity from near-infrared reflectance-patterns. This can be mathematically modelled to show behavior consistency so long as the ambient light is relatively consistent. This is because the frequency distribution drives the features, where a constant ambient is essentially a zero-frequency term. This presents a natural question: are the algorithms sensitive to noises that disrupt the image texture?

The primary noises relevant to this scope are camera and environment. These noises are common use-cases, but repeating the collection with them is challenging. It requires all the experiments to be redone while under noise presentation - a roughly 13-fold increase in imaging. One can imagine how this is both time consuming and expensive. Instead, a pragmatic simulation approach is proposed. Given the noise-physics is well understood, it is proposed they can be synthesized using semi-realistic generators.

The synthetic-augmentations are designed to realistically perturb the feature space. The noise generators are by design not photo-realistic. While using state-of-the-art simulation would likely generate more realistic noise, it would also require a precise CAD model of the faces (not available). This approach, however, does represent the actual physics at play. This should achieve the general desired results in a pragmatic fashion. The camera noise-augmentations are visualized in Fig. 5.2.

This spectroscopy algorithms sensitivity to noise is presented in Chapter-7. The noise-generation methods are detailed, explaining the motivation and implementation methodology (including the software tools). Algorithm sensitivity is then evaluated. This is first done using the existing algorithms on noise-augmented data, then includes evaluation of how noise-augmentations can be used as a training tool.

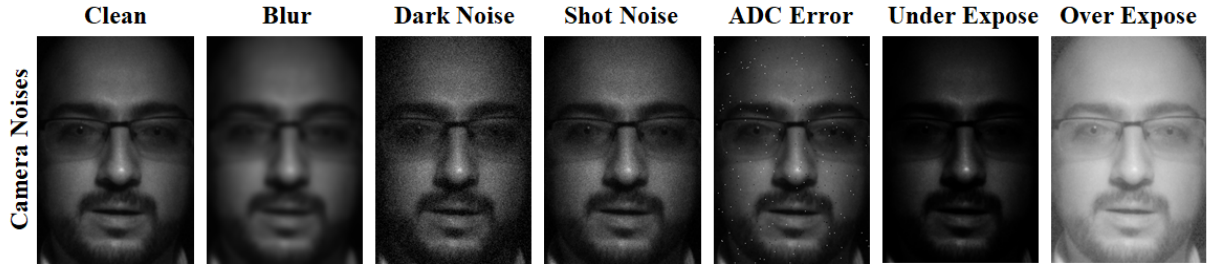


Figure 5.2: Introducing synthetic noise-augmentations. These sensor noise-generators are physics-informed but by design not photo-realistic. The intention is to perturb the features in a fashion that represents real-world noise.

5.3 Improving Spectroscopy Robustness with Auxiliary-Noise-Tasks

Deep-learning networks learn features from the classification task labels. Through loss back-propagation, the parameters are adjusted until they well differentiate the described classes (94). Sometimes intra-class variance, however, can be problematic. When classes overlap in labels, it gets difficult to identify the right features. Intuitively, one can infer that better class descriptions can improve the feature-set (where sometimes multiple labels are necessary).

This learning intuition can be applied towards noise robustness. The spectroscopy algorithms are sensitive to camera and environmental noises, where even training with noisy data results in a statistical degradation. This is clearly problematic. The noise does not actually change the liveliness; it only adds intra-class variance. Again, the goal is to be able to well describe this intra-class variance.

This variance sensitivity is addressed through a novel auxiliary-noise-task (ANT) framework. Following the intuition about label-precision, it is proposed that directly labelling noise presentations can help identify which features are associated with noise. These labels are then applied in a novel noise-inspired multi-task-learning application. This noise feature-encoding is theorized to ironically be a de-noising filter. By better distinguishing which features are associated with liveliness and which are associated with noise, optimizing classification layers.

Figure 5.3: Face crop with and without noise. Utilizing auxiliary-noise-tasks enables the network to learn which features are relevant to liveliness versus imaging noise.

An example of clean and noisy face-crops are visualised in Fig. 5.3. A traditional classification network runs the risk of being biased by the Gaussian noise. However, by jointly learning the primary classification and noise labels, the ANT framework can better isolate the right features.

This methodology is formalized in Chapter-8. A sensitivity analysis first is conducted to identify the optimal training method. While intuitively one can see benefits to jointly learning signal and noise features, actually learning orthogonal classification tasks can result in destructive parameter-interfere. A few novel loss paradigms are proposed to address this. Once the training method is optimized, the ANT networks are then benchmarked for liveliness performance. This learning methodology results in best-in-class robustness, and should be the go-to approach for monocular, single-frame FLV.

5.4 Image-Integrity-Verification via Camera-Noise

Image-integrity-verification (IIV) can be viewed as an extension of camera source-identification. Like a human fingerprint, images can be uniquely associated to their camera by noise characteristics (i.e. "noiseprints" (75; 73)). Intuitively, such a sensitive fingerprint should be impacted any time an image is tampered. This hypothesis has been pursued in a variety of fashions. Relevant examples for detecting noise tampering include discrete-wavelet-transform anomalies (78), color-filter-array anomalies (79), and photo-response-non-uniformity (PRNU) anomalies (77). PRNU arguably the most popular method to large-scale source-identification; intuitively, it should be sensitive to digital manipulations (19). The method of estimating PRNU is illustrated in Fig. 5.4.

Photo-realistic tampering can now be detected using DL based noise-analysis. The problem at hand is being able to do this detection in real-time. Current methods are complex, often requiring numerous frames (45; 84; 85). FR, conversely, requires an immediate response. Adding a notable delay for image-integrity-verification would fundamentally detract from the intended seamless experience.

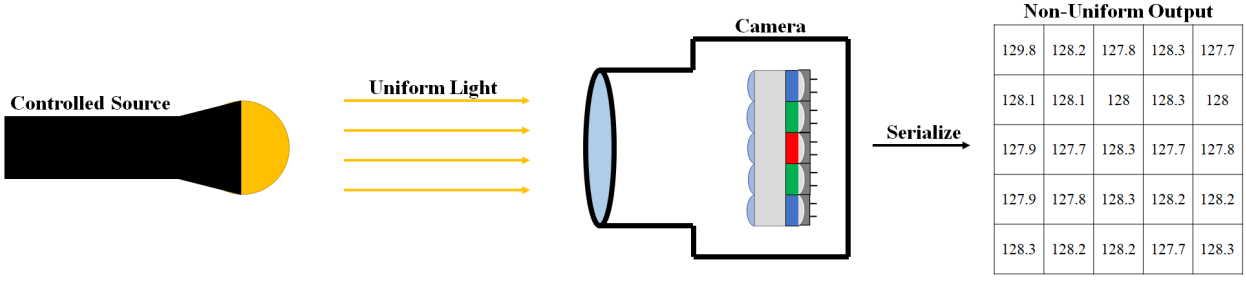


Figure 5.4: Photo-response-non-uniformity illustration. PRNU is the deviation in collected light from a uniform supply.

The proposed method is to verify image-integrity by comparing new frame PRNU values against an enrollment in a compressed. zonal fashion. This process is illustrated in Fig. 5.5. Anytime the new image value deviates sufficiently, the image is flagged as tampered. Conversely, small perturbations due to random noise must be tolerated as authentic. This is made possible through the zonal-analysis, which provides local sensitivity and global context. The compression then optimizes the latency for an imperceivable experience.

This proposal is possible because FR often utilizes an embedded camera. By owning the camera, it can be securely characterized for PRNU enrollment. This is formally presented in Chapter-9. The "noiseprint"-verification methodology is detailed, along with performance evaluation on face-swap-attacks and service-denial-attacks. The validation demonstrates FR can be secured against photo-realistic tampering in an extremely efficient fashion.

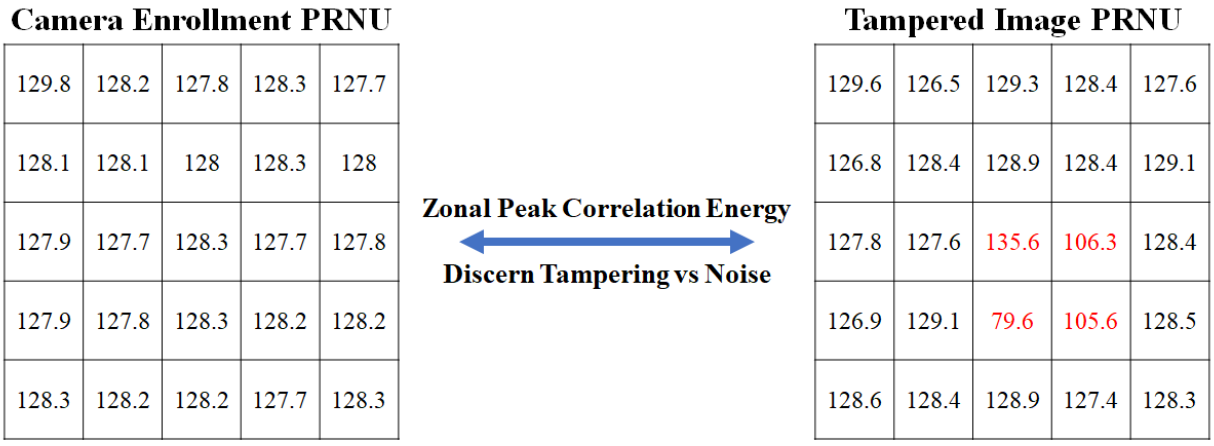


Figure 5.5: Verifying photo-response-non-uniformity for authenticity. This is a zone-based peak-correlation-energy comparison with the enrolled template.

5.5 Conclusion

In conclusion, four frameworks are proposed to addressing monocular, single-frame facial-liveliness-verification. Emphasis is placed on physical-spoof-attacks as the primary vulnerability. A general methodology is proposed (material-spectroscopy inspired) with two methods to improve noise robustness. Additionally, an image-integrity-verification framework is presented to mitigate face-swap and service-denial attacks. These secure the sensor against photo-realistic tampering. These four frameworks are evaluated in the next chapters.

CHAPTER 6

Monocular Facial-Liveliness-Verification via Near-Infrared Spectroscopy

The primary face-recognition (FR) vulnerability is the physical-spoof-attack (PSA). The National Institute of Standards and Technology (NIST) clarifies attack presentation methods in their ISO 30107 (14). These attacks are a physical "replay" of sorts, where a facsimile can spoof the FR algorithm into authenticating an imposter. Most spoofs are generated from 2D imagery and are easy to prepare. Examples include picture print outs, display photos or videos, and creating a mask from paper or fabric. These are highly effective and can be made from a headshot (acquired from social-media) and under $20 of materials. More complex methods can be done using 3D imagery, such as highly-realistic masks and mannequins. These methods, however, are generally unrealistic due to requiring a 3D model of the victim and therefore not in scope (1). For more details on the spoof presentations methods, see threat-model in Chapter-3.

The objective is to achieve robust facial-liveliness-verification (FLV) in monocular, single-frame fashion. 3D features are typically best-in-class for FLV. These can be generated using depth sensing (e.g., stereo vision, time-of-flight) (91), as well as multi-frame deep-learning networks (e.g., spatio-temporal) (92). While these methods generally robust, they do not meet the intended use-case. Hence, the proposal is to similarly learn material geometry features but in a monocular, single-frame fashion. This is achieved using a novel near-infrared (NIR) material-spectroscopy (MS) methodology.

Spectroscopy is an object identification technique that shines a controlled light source for reflectance-classification (93). This reflection is a function of the material's geometry and albedo. Intuitively, this means one can infer geometry from the reflectance-patterns as well. Live faces are a combination of multiple surfaces with varying radii of curvature. A mathematical model shows this results in a complex reflectance distribution, with varying frequencies as a function of radius of curvature. Conversely, the spoofing attacks considered here are either flat or essentially a simple convex surface. The mathematical model shows this results in a simple, low-frequency reflectance distribution. These distributions can be further biased by the material.

Figure 6.1: Juxtaposing the human and illuminated near-infrared perspectives. Note how illuminated near-infrared yields distinct reflectance-patterns. Live faces have complex texture; spoofs conversely appear washed out.

A visualization of these effects and how they can be used towards FLV is shown in Fig. 6.1. Note the differences in texture when comparing the liveliness presentations in the NIR perspective. The illumination helps discern the geometry, where it is clear the live face has a complex texture and the spoofs look washed out in comparison. This is expected behavior (a mathematical model formalizes this) and provides a notable advantage to liveliness. Spoofs that look realistic to the human eye now appear distorted and are much easier to detect.

Changing the spectrum also has secondary benefits of changing material-reflectivity. Attackers see in the visible band and naturally design facsimiles to realistic to their own eyes. The concept of "color," however, does not map linearly in the NIR spectrum. This can be also observed in Fig. 6.1, where the spoofs appear generally uniform (from using a single material) and live people demonstrate differences in the skin, eyes, hair, etc. This observation is not included in the mathematical model, but can help further differentiate the reflectance-patterns.

This chapter evaluates the performance of NIR MS under real-world operation conditions. A novel, large-scale data-set is collected to analyze live people and their corresponding spoofs under illuminated near-infrared imaging. This includes real-world noise factors, such as ambient lighting and changes in position and pose (approximately 80,000 unique frames). Robustness of the texture features is verified using a panel of 10 deterministic and deep-learning algorithms, demonstrating it is sufficiently reliable across numerous texture methods.

6.1 Facial-Liveliness Using Near-Infrared Spectroscopy

This chapter proposes a material-spectroscopy approach towards FLV. The hypothesis is as follows: facial-geometry can be characterized by shining NIR light and measuring the reflectance-patterns. This is modelled mathematically, where surface-reflectance equations are used to demonstrate there are necessarily differences between live and spoof faces. It is then proposed that texture methods can optimally do the FLV classification. Note the novelty is the near-infrared spectroscopy methodology - not a specific algorithm. Instead, the mathematical model is validated using a panel of texture algorithms. The expected reflectance distributions are illustrated in Fig. 6.2.

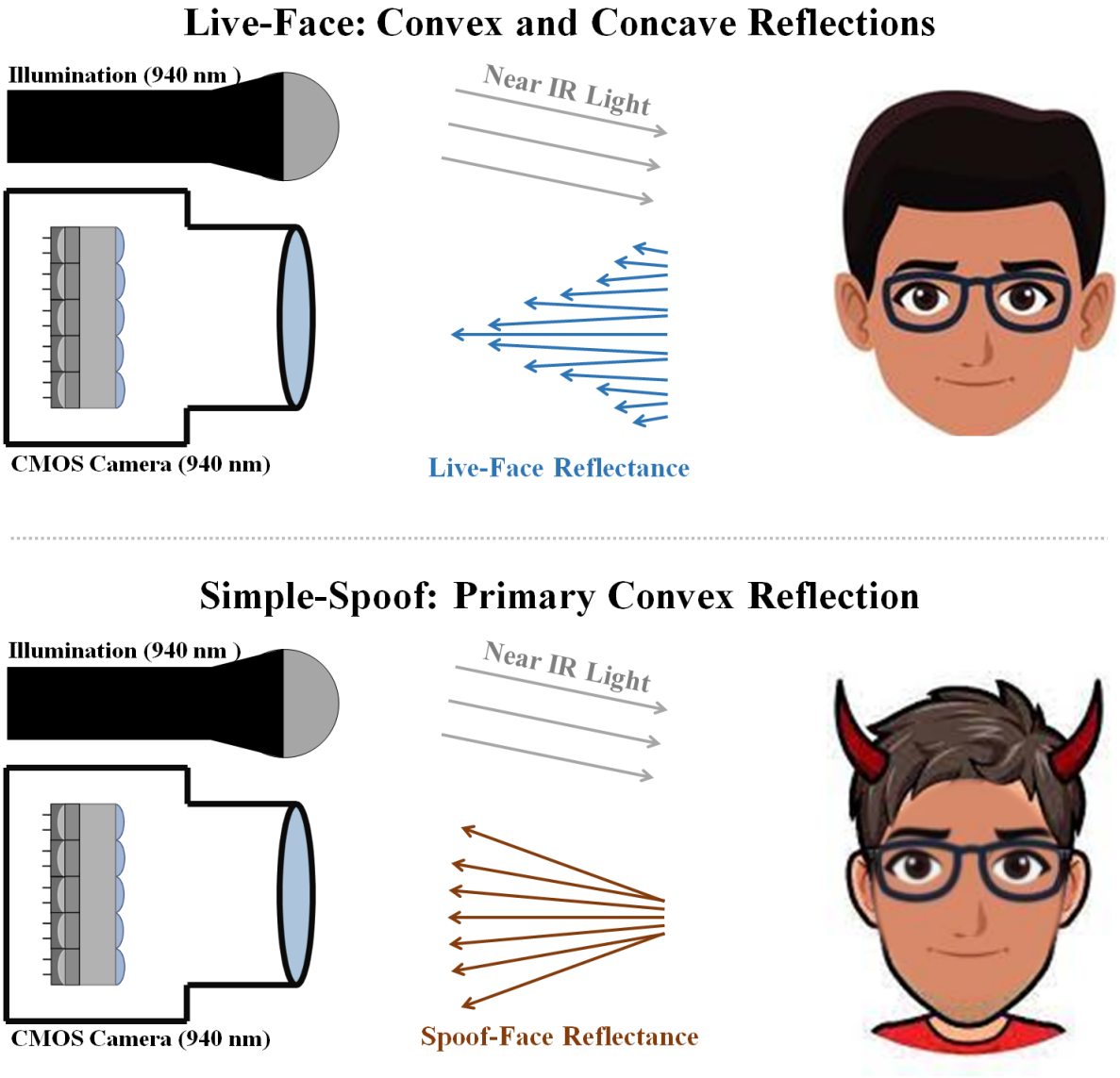


Figure 6.2: Facial reflectance-patterns: live versus simple-spoof. Live people have highly-variant reflectance patterns. Simple spoofs conversely have relatively uniform reflectance patterns.

Faces reflect the near-infrared illumination as a function of their geometry and albedo (reflectivity). Live faces have convex and concave components. This results in higher frequencies and varied reflectance-distributions (as a function of radii of curvature). Simple-spoofs, conversely, are generally a single convex-surface. This results in a simple, low-frequency reflectance-distribution.

Material albedo can also play a key role in classifying liveliness. Spoofs are designed to look realistic to the human eye; as such, the concept of "color" may not necessarily translate to other spectra. Recall how the spoofs look brighter and washed-out; this is because the commonly selected spoof materials, such as paper and fabric, are more reflective to NIR light than human skin (64). Note this reflectivity holds true over most NIR frequencies; 940 nm is selected because there is minimal contribution from solar radiance (95) or household LED light sources (96).

6.1.1 Facial-Reflectance Modelling

The facial-reflectance can be modelled by a combination of convex and concave Lambertian surfaces. A simple spoof, such as paper printout, is largely flat with slight convexity when bent to the face. This can be modeled as a simple convex sphere. A live face has convex and concave portions. For example, the nose is convex and the eye sockets are concave. This can be modeled as a primary sphere that has secondary convex and concave spheres. This research presents a proof that adding secondary convex and concave surfaces introduces additional frequency terms. These frequency terms can be used to infer geometric features and therefore determine liveliness. The recommend approach is to use texture classifiers.

For simplicity, this proof is done assuming only two-dimensions. By assuming a sphere, there is a uniform radius across all points; demonstrating the surface-reflectance on circles should intuitively hold true for spheres. This modelling is visualized in Fig. 6.3.

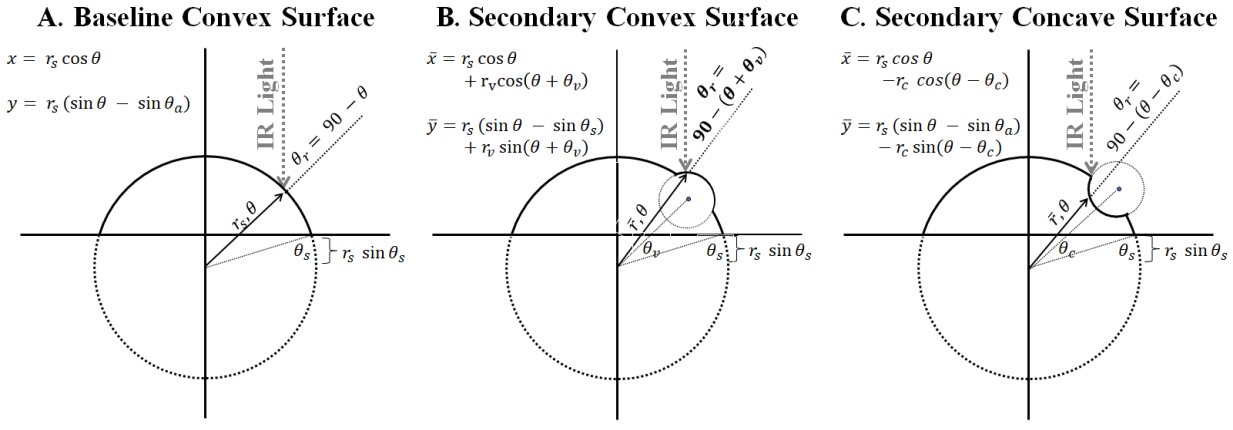


Figure 6.3: Facial-surface reflectance models. Model A is the baseline convex surface. Model B introduces a secondary convex surface. Model C introduces a secondary concave surface.

The three Lambertian surface-models are: simple-convex (A), secondary-convex (B) and secondary-concave (C). Note that these models are only relevant for the surface-angle range corresponding with positive y values; that is to say, θ is bound between the intercept angles θ_s and$\pi - \theta_s$. Furthermore, the models are valid only for the indicated type of surface. Model A is always valid, and Models B and C are only valid for angles bound within the secondary-surface (elsewhere defaulting back to Model A).

Lambertian surfaces reflect diffuse light as a function of the source intensity, angle from the surface normal and albedo (97). The surface reflection for an active illumination source is calculated in (6.1) . SR is the surface reflection, $\frac{\rho_d}{\phi}$ is the albedo (reflectivity coefficient), I is the infrared intensity and θ_r is the reflectance angle between incident ray and surface normal. Note that for now the facial-albedo is assumed to be a constant dependent associated with the primary material, such as skin, paper, glass, etc. This introduction of sub-components will be accounted for later.

$$SR = \frac{\rho_d}{\pi} I_r \times \cos(\theta_r) \tag{6.1}$$

The surface geometry directly impacts the infrared intensity. By definition, light waves decrease quadratically with the source-distance. This is generalized in (6.2) by assuming the I_0 is the light intensity, and the source-distance is difference between from the further part of the face, d_{face}, and surface height $y(\theta)$.

$$I_r = \frac{1}{4\pi} I_0 \times \frac{1}{(d_{face} - y(\theta))^2} \tag{6.2}$$

Calculating the surface height and normal vector can also be modeled as a function of surface-angle and radius of curvature. The surface angle θ is the angle with respect to the x-axis. The surface radius of curvature r_s is radius of the tangential circle. Note that the circle focal point may be offset on the y-axis. A perfectly circle surface by definition would have focal point at the origin. However, flatter surfaces would have a longer radius of curvature and therefore be an offset on the y-axis.

Let us assume that the surface angle which intersects the x-axis can be defined as θ_s. This means that the tangent circle focal point is then offset by $r_s \sin\theta_s$.

Recall the three Lambertian-surface models from Fig. 6.3. Model A is simplest, where the height is the x-axis projection and the surface-normal is the surface-angle. Introducing secondary surfaces, however, now it requires projecting a translation as a function of to the secondary surface-radius and angle to the secondary-arc focal-point. Model B presents a constructive interaction where the secondary convex-arc has focal-point at r_v, θ_v. The surface height is increased by the projected difference between the original arc and sub-arc focal-point, and the surface-normal is increased by θ_v in phase. Model C makes a similar assumption that the concave-arc focal-point is the

location r_c, θ_c. This time the translation is destructive; the secondary-surface height is decreased by the projected difference between the original arc and sub-arc focal-point, and surface-normal is decreased in phase by θ_c.

$$d(\theta) = \left\{ \begin{array}{ll} d_{face} - r_s \cdot (\sin(\theta) - \sin(\theta_s)), & \text{Model A} \\ \\ d_{face} - r_s \cdot (\sin(\theta) - \sin(\theta_s)) & \\ \quad -|r_s - r_v| \cdot \sin(\theta + \theta_v), & \text{Model B} \\ \\ d_{face} - r_s \cdot (\sin(\theta) - \sin(\theta_s)) & \\ \quad +|r_s - r_c| \cdot \sin(\theta - \theta_c), & \text{Model C} \end{array} \right\} \tag{6.3}$$

Surface-reflectance is a function of distance as applied in (6.3). More specifically, this distance is a function the furthest face-distance d_{face}, surface-radius r_s, surface-angle θ_s and surface bounding angle θ_s (noting secondary surfaces also have radii r_v, r_c with offset angles θ_v, θ_c). Intuitively, one can see the radius term, r_s, becomes dominant for flatter surfaces. This will inherently reduce the valid surface-angles to be closer to the y-axis (i.e., $\frac{\pi}{2}$), fundamentally acting as a low-pass filter. This is particularly relevant for 2D inspired spoofs, such as paper-mask and display-replay.

$$I(\theta) = SR(\theta) + I_{amb} \tag{6.4}$$

For completeness, the surface-reflectance also needs to factor the ambient light. This is simply a summation of the infrared surface-reflectance and light present at the surface I_{amb} as applied in (6.4). Ambient here is simplified to be diffuse. Note that the presence of image-noise is missing here. This is done to make the model solveable, but is necessary to analyze going forward. Sensor and environmental noises are realistic and will eventually effect FLV systems. This is addressed in the next chapter, presenting synthetic camera and environment noise-augmentations.

$$I(\theta) = \tfrac{\rho_d}{4\pi^2} I_0 \times \left\{ \begin{array}{ll} \frac{1}{(d_{face} - r_s(\sin(\theta) - \sin(\theta_s)))^2} & \text{Model A} \\ \quad \times \cos(\theta_I - \theta) + I_{amb}, & \\ \\ \frac{1}{(d_{face} - r_s(\sin(\theta) - \sin(\theta_s)) - |r_s - r_v| \sin(\theta + \theta_v)))^2} & \text{Model B} \\ \quad \times \cos(\theta_I - (\theta + \theta_v)) + I_{amb}, & \\ \\ \frac{1}{(d_{face} - r_s(\sin(\theta) - \sin(\theta_s)) + |r_s - r_c| \sin(\theta - \theta_c)))^2} & \text{Model C} \\ \quad \times \cos(\theta_I - (\theta - \theta_c)) + I_{amb}, & \end{array} \right\} \tag{6.5}$$

These equations can be combined for an explicit total reflectance equation. Adding secondary surfaces necessarily generates new frequency content in the reflectance profile as applied in (6.5). This in theory should generate more variance in the distribution - a behavior texture classifiers should easily identify. Note here the albedo, source intensity and ambient are constants and should not impact the light variance. Furthermore, the illumination source angle, θ_I, is visualized as 90 degrees in the surface-models figure but only acts as a phase shift. Position should theoretically matter less than the geometry of the object.

To verify the geometry is more important than position or light intensity, the surface-reflectance derivative can be taken. Complex surfaces have significantly more sinusoidal terms in the derivative (not shown for space reasons). This necessarily means there is a more complex variance across the surface intensity.

6.1.2 Liveliness Hypothesis

Putting these equations together yield a few interesting findings. First and foremost, the object flatness acts as a low pass filter. Recall how the surface angle is bound by the intersection angle θ_s; flatter objects have a longer radius of curvature and therefore a reduced range of surface angles. This is particularly relevant for simple spoofs, which can be largely flat by design (e.g., paper-mask or display-replay). Conversely, live faces are sufficiently curved to introduce a broader range of frequencies. This phenomenon is enhanced with complex geometries. Secondary convex and concave surfaces generate more variance in reflectance profile, with new inflection points caused by changing concavity (e.g., nose transitioning to eyes). Even fabric masks fail to fully capture these secondary surfaces. These can be thought of a low-pass filter on a real face, smoothing out the contours to behave like a single, convex surface. This behavior is indicated by (6.5) (and implied by the derivative, not shown), visualizing the different profiles.

6.1.3 Classification Methodology

Texture classifiers should be optimally situated to discriminate these reflectance profiles. In this regard, the methodology is not about a particular algorithm, but rather the material-spectroscopy process. This claim is verified in the experiment by benchmarking 10 relevant texture classifiers on the evaluation dataset. The goal is to demonstrate the spectroscopy approach generates very robust features, where the application developer can optimize performance and run-time.

6.2 Performance Evaluation

The experiments are designed to evaluate the material-spectroscopy approach to facial-liveliness-verification (FLV). Live actors and their corresponding spoofing attacks (paper-mask, spandex-mask, face-print, covid-mask and display-replay) are evaluated under various lighting and positions. These attacks are selected based off NIST ISO 30107 Levels A and B (14). The expectation is the combination of surface geometry and near-infrared albedo should be sufficient to translate liveliness into effectively a texture classification problem. To demonstrate this, a panel of deterministic and deep-learning texture algorithms are evaluated on the collected dataset.

Sample face-crops from the dataset are shown in Fig. 6.4. As described in the methodology, these spoofs have simple geometry and therefore reduced textural variance. Furthermore, spoof materials are generally more reflective, such that the spoofs often appear lighter. These effects can be well described by texture methods, including in the presence of ambient light. Lastly, note that the display-replay attacks do not show up to the NIR camera; this is essentially passive anti-spoofing.

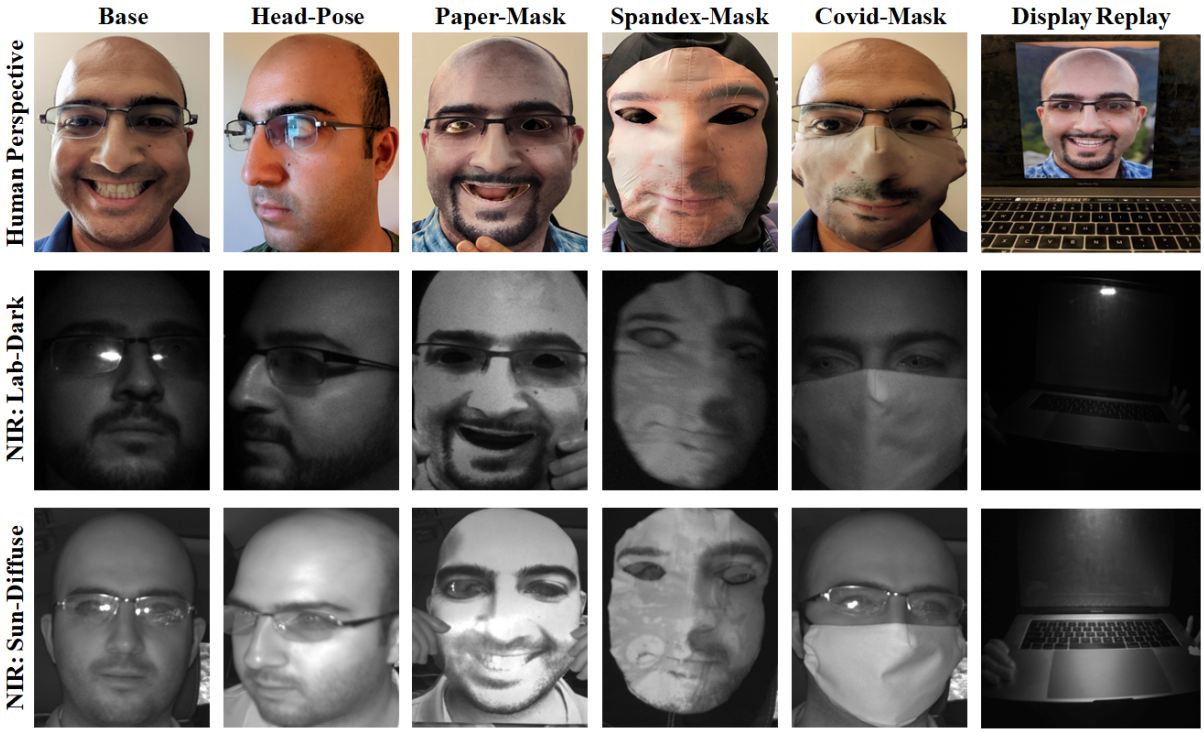


Figure 6.4: Data Collection Visualization. This matrix juxtaposes the different types of perspectives captured, noise and spoof attacks (noting that all spoofs also are present with varied pose and distance).

6.2.1 Texture Method Evaluation

To highlight the utility of infrared reflectance, a series of texture features are evaluated for FLV accuracy. The deterministic features are generated using texture descriptors and then classified using a random-forest-classifier (98) via science-kit-learn toolbox (99). The deep-learning features are classified using a fully-connected layer. All methods are evaluated using binary classification, live or spoof. Given some of the deterministic features do over-fit with the covid-mask spoof (which is hybrid of both classes), an ensemble is also evaluated for the 3D classes (deterministic features only). This ensemble is a majority vote for three binary classifiers trained to each spoofing-attack. Note that these methods are generally known for other texture classification applications. The novelty here is the application of illuminated near-infrared imaging for FLV via material-spectroscopy.

All algorithms are trained using stratified cross-validation. Test values are the average of ten randomized training and testing evaluations (80:20 participant ratio). Metrics include nominal-presentation-classification-error-rate (NPCER), attack-presentation-classification-error-rate (APCER) and average-classification-error-rate (ACER). These are essentially the false-rejection-rate of live people, the false-acceptance-rate of spoofs, and the average error rate. Note that ACER is selected over dataset accuracy as the data is imbalanced towards spoof presentations.

6.2.1.1 Local-Binary-Patterns

The first method is the local-binary-pattern (LBP) (100). The LBP are texture descriptors computed by a binary kernel. They are popular in facial-identification, and are demonstrated for liveliness by taking a multi-block approach (58). For FLV, the face-crop is first described with the LBP feature using the science-kit-image toolbox (101) and analyzed as a histogram feature (128 bins) using the science-kit-learn toolbox (99).

6.2.1.2 Discrete-Cosine-Transform

The second method is the discrete-cosine-transform (DCT) (102). The DCT is frequency analysis tool similar to the fast-Fourier-transform, but generally viewed as more explainable. It has historical uses in face-recognition (103) and can theoretically discriminate the liveliness classes by reflectance frequency content. For FLV, the face-crop is first described by the DCT algorithm using the science-python toolbox (104), where the frequency terms are extracted via zig-zag pattern for classification. Principle-component-analysis (PCA) is used to reduce the feature space to 128.

6.2.1.3 Ranked Channel Histograms

The third method is the Ranked-channel-histogramss (RCHs) (105). Others have shown that the distribution of color channels can be used for liveliness, where they used principal component analysis to identify the key components from each channel's histogram. This approach can similarly be used to describe the variance of infrared light reflection, and therefore correlate with the texture. For FLV, the face-crop is described by its histogram using the science-kit-learn toolbox (99) and then the 25 primary bins as determined by PCA are used for classification.

6.2.1.4 Random-Fourier-Series

The fourth method is the random-Fourier-series (RFS), also known as random kitchen sinks (106). This feature is theoretically similar to the DCT; however, instead of explicitly calculating relevant frequency features, they are generated as a sum of randomized series. Prior applications have shown promise using random features, though there is risk a poor set of frequency series are selected. For FLV, The face-crop is first described by the RFS algorithm using the science-kit-image toolbox (101), where the output is then unraveled into an array for classification. Principal-component-analysis is used to reduce the feature space to 128.

6.2.1.5 Deep-Learning

The final approach is to describe the face using deep-learning (DL), implicitly describing the texture as a function of the convolutional layers (107). For FLV, the face-crop is described by a convolutional-neural-network (CNN) feature encoder then classified using a 128-neuron feature fully-connected layer and 2-neuron classification layer. CNN features are evaluated by employing the MobileNetV2 encoder (108) for an efficient network and the InceptionV3 encoder (109) for a robust network. Both encoders are pre-trained on ImageNet (110) and fine-tuned on the evaluation dataset. Training and evaluation are conducted using the Tensorflow toolbox (111).

6.2.2 Research Limitations

Facial-spoofing is inherently an evolving field, where each counter-measure will eventually be exposed by attackers. The attack presentation methods selected are designed to represent the most common attacks with relevant noises, but are inherently not inclusive. A general pragmatic approach to algorithm development would be to verify it meets requirements on this evaluation dataset, then make updates based upon penetration testing.

6.2.3 Exp 1: Laboratory Dark and Diffuse-Light Liveliness Evaluation

The first experiment is recorded a camera laboratory. The general concept is to image the key liveliness perspectives in a controlled setting. This dataset is controlled of 30 diverse adults. Gender is represented by 20 males and 10 females. Ethnicity is represented by 11 Caucasian, 8 Asian-Southeast, 3 Asian-Pacific, 3 Middle-Eastern, 3 Hispanic and 2 African participants. Age is presented by 4 in 18-24 range, 11 in 25-34 range, 2 in 35-44 range, 8 in 45-54 range and 6 in 55-64 range. Each participant is coached to perform the same procedure for fully-contrastive imaging.

Each participant starts with sitting in the camera lab, looking at the camera and performing several head nodding motions. These perspectives include distance ranging from .5 to 1.5 meters, yaw ranging from -45 to 45 degrees and pitch ranging from -15 to 15 degrees. This process is performed under laboratory dark conditions (940 nm illumination only) and laboratory light conditions (940 nm illumination with all lab lights turned on). This process is then repeated for the spoofing attacks: display-replay, paper-mask, spandex-mask and customized covid-mask.

The dataset is next pre-processed for algorithm training. All images segmented by face-crop using the RetinaFace algorithm (12). Face-crops are then scaled to 225 x 225 for deterministic algorithms or the associated DL network-input size as appropriate. No further adjustments are made to keep the image-quality representative of real-world conditions.

These liveliness presentations are given in Table 6.1. This collection is a robust contribution both due to its spectral content and strong variance in attributes. There are approximately 62,000 fully-contrastive laboratory images; all participants have their live and spoof presentations under the varied perspectives. Imaging is done with a 5 mega-pixel FLIR Blackfly monochrome camera (112) employing a 940 nm filter with matching illumination source. Note that 940 nm is selected because of the minimal solar contributions (95) (minimizing interference).

Presentation	**Ambient**	**Distance (meters)**	**Yaw (deg)**	**Pitch (deg)**
Live (30)	Dark, Lights	[.5, 1.5]	[-45, 45]	[-15, 15]
Display-Replay (30)	Dark, Lights	[.5, 1.5]	[-45, 45]	[-15, 15]
Paper-Mask (30)	Dark, Lights	[.5, 1.5]	[-45, 45]	[-15, 15]
Spandex-Mask (30)	Dark, Lights	[.5, 1.5]	[-45, 45]	[-15, 15]
Face-Print Covid-Mask (30)	Dark, Lights	[.5, 1.5]	[-45, 45]	[-15, 15]

Table 6.1: Near-infrared material-spectroscopy liveliness presentation matrix: laboratory conditions. All participants are imaged under every combination in the matrix.

6.2.3.1 Exp 1: Results

The NIR MS methodology proves to be rather robust under laboratory conditions. The results given in Table 6.2 (see next page) verify the mathematical model. The expectation is live faces should necessarily have more high frequency content and potential inflections where concavity changes. These behaviors are correctly observed by the texture classifiers. Most algorithms correctly detect the presence of spoofs (APCER metric) without risk of falsely-rejecting live people (NPCER metric). It is particularly promising that the 3D spoofing classes are easily detected; this is likely due to the materials being extremely reflective. Note that the display-replay attacks are all represented as not-detected (ND) as no faces are observed. These are not reflected in the classification rates.

This demonstrates that a robust texture descriptor can differentiate the liveliness materials. The local-binary-pattern classifiers are particularly robust, where there is both strong average precision and minimal bias towards any specific class. The discrete-cosine-transform and ranked-channel-histograms similarly show good general robustness; however, it appears the covid-mask class causes some bias towards false-rejections. This bias can be addressed through the use of the 3D ensemble, which uses the average performance across classifiers tailored to each attack vector. Note how not only the average precision improves, but the error rates are better distributed. The random-Fourier-series however seem to be inconsistent. This is reasonable as even with principal-component-analysis the features are selected at random.

Algorithm	**ACER**	**NPCER**	**APCER:**			
			Paper	**Spandex**	**Covid**	**Display**
LBP †	3.6%	4.9%	6.4%	0.2%	0.0%	ND
DCT	9.8%	17.6%	1.0%	5.2%	0.0%	ND
RCH	14.9%	23.3%	11.5%	4.6%	0.0%	ND
RFS	50.1%	99.9%	0.4%	0.1%	0.1%	ND
LBP - 3DE	1.7%	0.9%	7.7%	0.0%	0.0%	ND
DCT - 3DE	3.8%	3.8%	2.2%	8.4%	0.5%	ND
RCH - 3DE	6.7%	8.1%	10.2%	0.4%	5.5%	ND
RFS - 3DE	24.8%	31.0%	17.0%	17.0%	22.0%	ND
MobileNetV2 ‡	1.0%	1.1%	2.5%	0.0%	0.0%	ND
InceptionNetV3	0.3%	0.4%	0.5%	0.0%	0.0%	ND

Table 6.2: Near-infrared material-spectroscopy liveliness results: laboratory conditions. † indicates optimal deterministic algorithm. ‡ indicates optimal deep-learning algorithm.

Both deep-learning networks perform exceptionally. The InceptionV3 (109) network has near perfect accuracy across all classes, showing strong feature robustness. Perhaps more interesting is the MobileNetV2 (108) network is essentially just as robust. This verifies a small feature space is sufficient for the texture classification. It is believed the key to this robustness is the contrastive dataset. Recall that all participants are imaged under the same procedure for all liveliness presentations. This is theorized to assist the networks in discerning relevant liveliness features from random noise.

6.2.4 Exp 2: Exterior Sun-Load Liveliness Evaluation

Experiment 2 extends the data-collection to exterior conditions. The first experiment is controlled as to verify the mathematical model; this is clearly successful. The mathematical model also implies that ambient light should not present any texture-noise when it is (relatively) uniform. This is evaluated next by repeating the experiment under outdoor sun-load. This helps validate the MS methodology robustness under real-world conditions.

For simplicity, the sun is introduced under diffuse conditions. The participants are placed in a vehicle such that the light is diffused by glass. Introducing specular sources that directly bias the camera is a hard use-case for even well-known applications. Hence, this chapter does a simple sun-load investigation, and the next chapter introduces specular sources.

The additional exterior collected conditions are given in Table 6.3. This is approximately an additional 18,000 images, resulting in a total of 80,000 unique presentations. Note that the sun-diffuse condition is a retro-active add-on to the laboratory experiment; all spoofing attacks are presented but only a subset of the original participants are available for further imaging. Furthermore, the procedure is optimized, allowing for fewer frames to sufficiently capture the position and pose use-cases.

Presentation	**Ambient**	**Distance (meters)**	**Yaw (deg)**	**Pitch (deg)**
Live (30)	Dark, Lights, Sun*	[.5, 1.5]	[-45, 45]	[-15, 15]
Display-Replay (30)	Dark, Lights, Sun*	[.5, 1.5]	[-45, 45]	[-15, 15]
Paper-Mask (30)	Dark, Lights, Sun*	[.5, 1.5]	[-45, 45]	[-15, 15]
Spandex-Mask (30)	Dark, Lights, Sun*	[.5, 1.5]	[-45, 45]	[-15, 15]
Face-Print Covid-Mask (30)	Dark, Lights, Sun*	[.5, 1.5]	[-45, 45]	[-15, 15]

Table 6.3: Near-infrared material-spectroscopy liveliness presentation matrix: laboratory and exterior conditions. Exterior diffuse-sun (indicated by *) has all spoof combinations with a subset of live participants; all others are fully contrastive.

6.2.4.1 Exp 2: Results

The exterior evaluation results given in Table 6.4 verifies the methodology is robust. This conceptually makes sense. The mathematical model indicates the diffuse ambient-light is essentially a zero-frequency term and thus does not affect texture features. The only real concern is whether the images could be saturated by light, which is addressed by the camera's auto-exposure functionality.

Algorithm results are generally within statistical noise of the laboratory conditions, sometimes actually improving. This is believed to be a result of the diffuse sun-load adding constructively with the illuminator. Perhaps most interesting is the ranked-channel-histograms remain stable across lighting conditions. This particular feature is hypothesized to be the most light-sensitive, but the relative distribution remains sufficient even without the precision of a texture kernel.

Algorithm	**ACER**	**NPCER**	**APCER:**			
			Paper	**Spandex**	**Covid**	**Display**
LBP †	2.6%	4.0%	3.5%	0.1%	0.1%	ND
DCT	7.9%	13.5%	1.2%	5.6%	0.0%	ND
RCH	9.4%	15.9%	8.3%	0.0%	0.1%	ND
RFS	49.7%	7.5%	91.4%	96.8%	87.6%	ND
LBP - 3DE	1.6%	2.4%	1.9%	0.2%	0.1%	ND
DCT - 3DE	4.7%	2.9%	5.5%	13.3%	0.4%	ND
RCH - 3DE	6.7%	10.6 %	8.0%	0.4%	0.3%	ND
RFS - 3DE	50.0%	99.9%	0.1%	0.2%	0.1%	ND
MobileNetV2 ‡	0.9%	0.9%	2.6%	0.0%	0.0%	ND
InceptionNetV3	0.4%	0.4%	0.7%	0.2%	0.0%	ND

Table 6.4: Near-infrared material-spectroscopy liveliness results: laboratory and exterior conditions. † indicates optimal deterministic algorithm. ‡ indicates optimal deep-learning algorithm.

6.3 Conclusions

Spoofing attacks are a major vulnerability for current face-recognition systems (14). With minimal effort, attackers can make realistic facsimiles for authentication. Given most spoofs are inspired from 2D images, depth technologies are historically used. As such, this research a novel near-infrared material-spectroscopy approach to achieve monocular, single-frame facial-liveliness-verification.

The spectroscopy methodology employs geometric-features by illuminating faces with near-infrared light and analyzing the reflectance-patterns. This approach has a secondary benefit from using a non-visible spectrum, which introduces differences in material-reflectivity. This approach results in extremely robust experimental results.

Depending on the available hardware, either the local-binary-pattern (LBP) or MobileNetV2 algorithms are recommended. If it is a CPU based platform, the LBP algorithm is sufficiently robust and can be done in real time. If a deep-learning accelerator is available, the MobileNetV2 algorithm is recommended. It is extremely robust and can be run with negligible latency to the user. Note that while the 3D ensemble can address bias from the covid-mask classes (which are essentially a hybrid of live and spoof), it generally is not worth the extra compute. The optimal algorithms (LBP or MobileNetV2) are robust enough to not need the ensemble, whereas the others at best match their robustness when employing the ensemble.

A necessary next step is to expand the dataset noise factors. This collection is designed to verify the proof, meaning the ambient light is structured to be diffuse. It is important to verify robustness against real-world image-noises, such as camera and environmental artifacts. These are evaluated in the next chapter.

CHAPTER 7

Addressing Noise via Synthetic Generators

Near-infrared (NIR) material-spectroscopy (MS) is a robust means of facial-liveliness-verification (FLV). Liveliness is deteremined by illuminating faces with NIR light and classifying reflectance-patterns. These reflections are a function of geometry and material-reflectivity. The live faces should have complex frequency distributions, whereas the spoof faces should comparatively only have simple low-frequency content. This hypothesis is validated through both mathematical modelling and empirical evaluation in Chapter-6.

One of the shortcomings of the NIR MS evaluation is the lack of image-noise. While there is a lot of variance in the scenarios (liveliness-presentation, facial-distance, facial-pose and ambient-light), the dataset is recorded with a high-quality camera under diffuse light conditions. While this evaluation successfully verifies the mathematical model (the desired intent), real-world face-recognition (FR) applications are expected to work in the presence of image-noise.

This yields a couple of fundamental questions. First and foremost: are the texture algorithms sensitive to camera and environmental noises? If the algorithms are sensitive, a second question naturally follows: can noisy training data improve overall performance? There is potential risk training on noisy data could over-fit to the noises, and degrade the classification accuracy on clean data.

This chapter proposes a synthetic noise-generation framework to address these in a cost-effective fashion. Repeating the entire experiment with camera and environmental noises is time-consuming and expensive. These noises, however, have well understood physics. For example, photo-receptors leak current under low light (dark noise, Gaussian distribution) and have uneven acquisition at bright light (shot noise, Poisson distribution). These physics models can be translated into semi-realistic generator-algorithms for augmenting the existing dataset. This process is done to evaluate the optimal spectroscopy algorithms for noise-sensitivity. This is first conducted by only using the existing dataset for training, then exploring how noise-augmentations can be used as training tools.

7.1 Designing Semi-Realistic Noise Generators

The goal at hand is to design semi-realistic camera and environmental noise generators. A semi-realistic noise generator is one that augments the image in a fashion that represents real-world physics, but does not necessarily look photo-realistic. The idea here is to perturb the feature space in physics-informed fashion, acting as a desensitization tool (e.g., a vaccine). While a simulation-to-real approach would result in better image quality, this requires a 3D model for each participant and their spoofs (which is not available).

Camera and environmental noises are analyzed as they are most relevant to FR algorithms. For example, it is possible for images to corrupt in storage or memory, but these are failures with the hardware that are independent of the software. All sensors, however, are likely to have some noise caused by the sensor and auto-exposure algorithms. Furthermore, it is inevitable that the sun will eventually introduce bright spots and shadows on the user's face. These noises are extremely difficult to filter out with hardware. Given their frequency of occurrence, it seems logical to address algorithm robustness to them.

The relevant noise factors identified and their generation tools are given in Table-7.1. These are selected because they meet two criteria: the noise is very likely to present in real FR use-cases and the associated physics can be simulated with computer-vision tools. The expected effects and generation methods are detailed for the selected noises next.

Noise	**Generator**
Camera focus[1] (blurriness)	Gaussian Low Pass Filter
Dark noise[1] (random leakage)	Gaussian Noise Generator
Shot noise[1] (random photo distribution)	Poisson Noise Generator
Salt and pepper noise[1] (analog-to-digital error)	Random 0 and 255 Generator
Under-exposure[1] (low contrast)	Gamma Subtraction
Over-exposure[1] (saturation)	Gamma Addition
Point-source[2] (point sources)	Synthetic Bright Ellipse
Point-shadow[2] (shadows)	Synthetic Dark Ellipse
Streaking-source[2]	Synthetic Overhead Sun
Streaking-shadow[2]	Synthetic Overhead Shadow
Piping-source[2]	Synthetic Side Sun
Piping-shadow[2]	Synthetic Side Shadow

Table 7.1: Synthetic camera and environmental noise augmentation generators. This table enumerates both the relevant types of noises and how they are being generated for evaluation. Camera noises are indicated by [1]. Environmental noises are indicated by [2].

7.1.1 Camera Noise: Focus

Camera focus is required to ensure a crisp facial-image. A lens being out of focus consequently then results in blurry imagery (113). This is simulated using a Gaussian blurring kernel. The noise generator is implemented utilizing the Science-Kit Image toolbox (101).

7.1.2 Camera Noise: Dark-Current

Camera photo-receptors are imperfect, and can leak current even when no light is supplied (114). This effect is essentially randomly supplying pixel intensities, and can simulated using a Gaussian distribution. The noise generator is implemented utilizing the Science-Kit Image toolbox (101).

7.1.3 Camera Noise: Shot

Light is really ever perfectly uniform. The photons are often received in stochastic process, which is defined as shot noise (115). This effect can be modelled by using a Poisson process, effectively adding a "pepper" effect to the image. The noise generator is implemented utilizing the Science-Kit Image toolbox (101).

7.1.4 Camera Noise: Under-Exposure

Cameras typically expose light until a basic set of contrast metrics are met (116). These metrics are often a simple count of white and black pixels, and as such the overall image can be under-exposed if there are notable bright spots in the scene (116). This causes the exposure-time to get biased incorrectly to be too short, resulting in the face being very dark (non-coincidentally also presenting dark-current noise). This effect can be modelled by adjusting the gamma to be darker such that facial-features start to disappear. The noise generator is implemented utilizing the Science-Kit Image toolbox (101).

7.1.5 Camera Noise: Over-Exposure

Over-exposure is the opposite problem of under-exposure. Due to dark spots in the image, the overall exposure time is increased to be too high. This results in the face looking saturated (116). This effect can be modelled by adjusting the gamma to be brighter such that facial-features start to disappear. The noise generator is implemented utilizing the Science-Kit Image toolbox (101).

7.1.6 Environment Noise: Point-Source

Point sources present in a point-like fashion on the face (117). This results in the region being particularly bright, often presenting shot noise (with non-source region consequently under-exposed). This is simulated using a randomized ellipse using OpenCV (118). The region within the ellipse is then over-exposed and the region outside the ellipse is under-exposed (using the generators proposed). The boundary between regions is blurred.

7.1.7 Environment Noise: Point-Shadow

Point shadows are the inverse of a point source (117). This results in the shadow region being particularly dark, often presenting dark-current noise (with the non-shadow region consequently over-exposed). This is simulated using a randomized ellipse using OpenCV (118). The region within the ellipse is then under-exposed and the region outside the ellipse is over-exposed (using the generators proposed). The boundary between regions is blurred.

7.1.8 Environment Noise: Streaking-Source

In some cases, a specular source may present itself overhead the user. If they have an obstruction, such as wearing a hat or the roof of a vehicle, anecdotal evidence shows this results in the bottom half of the face being illuminated by a bright streak (with non-streak region consequently under-exposed). This is simulated using a randomized streak using OpenCV (118). The region within the streak is then over-exposed and the non-streak region is under-exposed (using the generators proposed). The boundary between regions is blurred.

7.1.9 Environment Noise: Streaking-Shadow

Opposite to a light streak, the specular source may present itself below the user. This can happen when the user is in an elevated position, such as riding in a vehicle. Anecdotal evidence shows this results in the top half of the face being illuminated by a bright streak (with the non-streak region consequently under-exposed). This is simulated using a randomized streak using OpenCV (118). The region within the streak is then over-exposed and the non-streak region is under-exposed (using the generators proposed). The boundary between regions is blurred.

7.1.10 Environment Noise: Piping-Source

Alternatively, the specular source may present at angle too the user. This can happen when the user is facing north or south, and the sun is oriented to the east or west (depending on time of day).

Anecdotal evidence shows this creates a bright, light-pipe across the user's face (with the non-pipe region consequently under-exposed). This is simulated using a randomized pipe across the face using OpenCV (118). The region within the pipe is then over-exposed and the non-pipe regions are under-exposed (using the generators proposed). The boundary between regions is blurred.

7.1.11 Environment Noise: Piping-Shadow

Lastly, a specular source can be obstructed by a large object that casts a piping shadow. This can occur when the user is underneath a large structure, such as driving under a bridge. Anecdotal evidence shows this creates a dark, shadow-pipe across the user's face (with the non-pipe region consequently over-exposed). This is simulated using a randomized pipe across the face using OpenCV (118). The region within the pipe is then under-exposed and the non-pipe regions are over-exposed (using the generators proposed). The boundary between regions is blurred.

7.2 Performance Evaluation

These experiments are designed to evaluate the material-spectroscopy sensitivity to semi-realistic noises. First, algorithm sensitivity to camera and environmental noise is evaluated (e.g., train on clean, evaluate on noise). Next, the utility of using noise-augmentations as a training tool is explored. Examples of the noise-augmentations for each generator are visualized in Fig. 7.1.

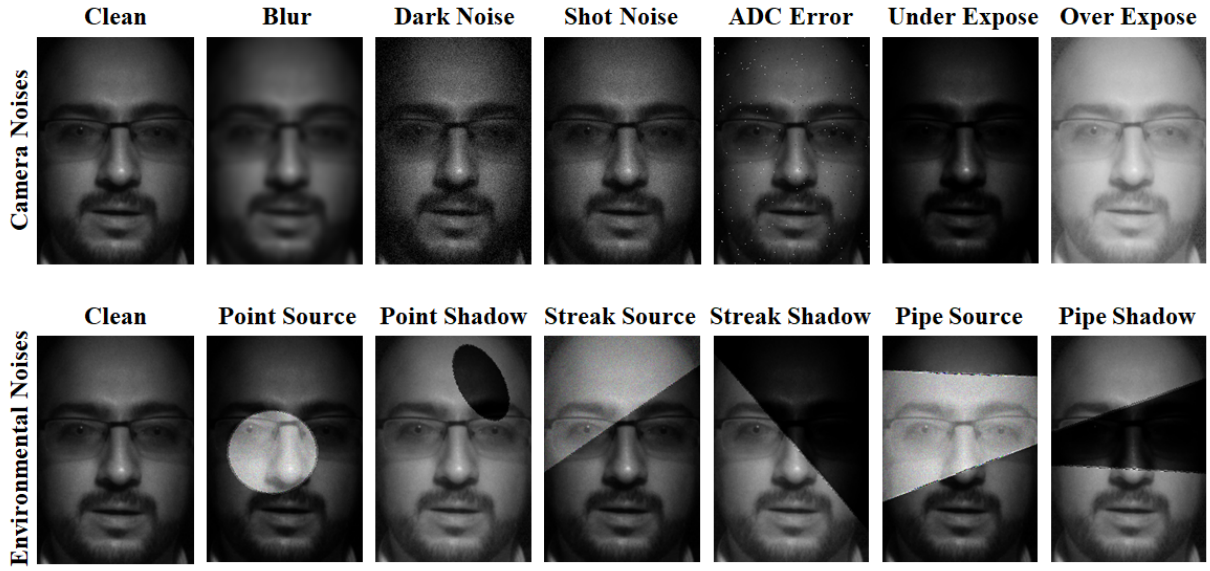


Figure 7.1: Visualizing synthetic camera and environment noise-augmentations. First row are the camera noises; second row are the environmental noises.

All evaluations incorporate stratified cross-validation. Each experiment is evaluated 10 times, randomizing the training and testing participants to elucidate algorithm sensitivity. The evaluation metrics are nominal-presentation-classification-error-rate (NPCER), attack-presentation-classification-error-rate (APCER) and average-classification-error-rate (ACER). These are essentially the rate at which live faces are mis-classified, the rate at which spoofs are mis-classified and the average of both respectively. Final reported scores are the average of the 10 stratified-cross-validation testing values.

7.2.1 Validation Algorithms

Given this is a data augmentation methodology, the validation algorithms are the optimal deterministic and deep-learning algorithms from the NIR MS study. The optimal deterministic algorithm is the local-binary-pattern (LBP), achieving 2.6% test ACER. The optimal deep-learning algorithm is the MobileNetV2, achieving a rather excellent 0.9% test ACER. Note that the InceptionV3 network is incrementally more accurate; however, this small benefit is not viewed to be worth the computational-costs.

7.2.2 Research Limitations

The noises evaluated here are synthetically generated from physics-inspirations. By definition, they are not going to be as realistic as a simulation that incorporates a 3D model of the participants. Furthermore, these noises are also a selected subset based upon which have sufficiently known physics models. If one additional noise is to be modelled, rain is likely the best choice for its prevalence. Rain is not as well understood as the noises modelled; however, intuitively one can imagine it to be a series of randomized refracting blobs. This could also be simplified to just blurring blobs. Future works may investigate this.

7.2.3 Exp 1: Synthetic-Noise Sensitivity

The first experiment is to evaluate the spectroscopy algorithms sensitivity to the camera and environmental noises. The algorithms are all trained on the clean (non-augmented) data, then evaluated on the noise-augmented dataset. Test evaluation is also done on clean data as a control. Noise-augmentation generators are evaluated as just camera, just environment and both. Note that evaluating both is effectively redundant as the generators are intentionally not combined, but still useful to see as the real-world use-case.

Algorithm	Train	Test	ACER	NPCER	APCER
LBP	Clean	Clean	2.6%	4.0%	1.2%
LBP	Clean	Cam Noise	29.0%	27.9%	30.1%
LBP	Clean	Env Noise	37.8%	65.6%	10.0%
LBP	Clean	Cam & Env Noise	33.1%	44.0%	27.9%
MobileNetV2	Clean	Clean	0.6%	0.3%	0.9%
MobileNetV2	Clean	Cam Noise	13.7 %	11.5%	16.0%
MobileNetV2	Clean	Env Noise	18.2 %	17.9%	18.5%
MobileNetV2	Clean	Cam & Env Noise	33.0%	0.9%	0.9%

Table 7.2: Evaluating near-infrared material-spectroscopy sensitivity to synthetic camera and environmental noise. Camera is abbreviated to Cam. Environment is abbreviated to Env.

7.2.3.1 Exp 1: Results

The experimental results unfortunately show the NIR MS algorithms are sensitive to both camera and environmental noise. Regardless of whether deterministic or deep-learning, algorithms trained on clean data do not do well on camera or environmental noises. Furthermore, the algorithms are more sensitive to environmental noises versus camera. This is intuitively reasonable. The environmental noises are designed to potentially occlude the faces, whereas many of the camera noises simply degrade the image sharpness. These results are shown in Table 7.2.

7.2.4 Exp 2: Augmenting Training with Synthetic Noise

The first experiment shows both the deterministic and deep-learning NIR MS algorithms are sensitive to camera and environmental noises. The natural next step is to investigate the impacts of including noise in the training set. The goal is to improve robustness to noise without degrading performance on clean data.

A similar experiment is repeated, but this time introducing noise in the training dataset. When noise-augmenting, 10% of the data is randomly selected for noise-augmentation (both camera and environmental). This value is selected to keep approximately a 1-1 ratio of clean and noisy images. When testing, all images are noise-augmented by all generators for completeness. The test evaluation is conducted on both clean and noise-augmented data with two controls: only-clean training and only-noise-augmented training. The purpose of only training on noise-augmented data is to understand if the classifiers can intuitively separate noise from the liveliness signal.

Algorithm	Train	Test	ACER	NPCER	APCER
LBP	Clean	Clean	2.6%	4.0%	1.2%
LBP	Clean	Noise	33.1%	44.0%	27.9%
LBP	Noise	Clean	2.8%	4.7%	1.0%
LBP	Noise	Noise	4.5%	7.6%	1.4%
LBP †	Clean & Noise	Clean	2.8%	4.6%	1.0%
LBP †	Clean & Noise	Noise	4.5%	7.6%	1.4%
MobileNetV2	Clean	Clean	0.6%	0.3%	0.2%
MobileNetV2	Clean	Noise	18.2%	17.9%	18.5%
MobileNetV2	Noise	Clean	0.7%	0.4%	0.9%
MobileNetV2	Noise	Noise	1.7%	0.5%	3.0%
MobileNetV2 †	Clean & Noise	Clean	0.7%	0.7%	0.7%
MobileNetV2 †	Clean & Noise	Noise	1.0%	0.6%	1.4%

Table 7.3: Evaluating near-infrared material-spectroscopy with synthetic noise-augmentation training. Algorithms are trained using 10% camera and environmental noise-augmentations (randomly distributed), then evaluated on clean and noisy data (note noisy-test is 100% noise-augmented). Noise here implies both camera and environmental noise generators. Algorithms trained with the optimal data composition are indicated by the †.

7.2.4.1 Exp 2: Results

The results given in Table 7.3 are far more promising. The optimal training method is clearly to include clean data and noise-augmentations (indicated by the †). This dramatically improves test performance on noisy data with negligible degradation on clean data. This phenomenon is exhibited by both the LBP and MobileNetV2 algorithms.

These findings indicate noise-augmentation should absolutely be used. This addresses the sensitivity concern without degrading clean performance, implying there is zero risk - only benefits. Perhaps more interesting is that training only noise-augmented data is inline with combining clean and noise-augmented data. This suggests that the classifiers are able to implicitly identify liveliness signal from synthetic noise when training.

7.2.5 Exp 3: Improving Training-Data Contrast with Synthetic-Noise

These results are achieved by using a full-contrastive dataset; i.e., every participant is imaged under every condition. This collection method is very robust and ideal for training. However, it is also rather time consuming and expensive. The second experiment results raise an interesting question: can noise-augmented data be used as a replacement for collecting additional clean data?

This experiment is an evaluation of data degradation sensitivity. The experimental design is repeated (i.e., same training and testing methods), but the participant data is intentionally degraded. Two new contrast-degraded states are introduced: partially-contrastive and not-contrastive. The partially-contrastive dataset randomly selects between live, spoof or live and spoof presentations for training. That is to say all the data exists, but only a subset of liveliness classes are selected to ensure roughly one-third are fully contrastive. The not-contrastive dataset randomly selects between live or spoof presentations for training - but no participants have all presentations.

Intuitively, the partially-contrastive dataset should converge but potentially be less accurate. This is because there are fully contrastive participants to best separate liveliness classes, and the other participants can simply re-enforce these features. The not-contrastive, however, is not necessarily guaranteed to converge. Without any contrastive pairs, there is risk this transforms from a liveliness problem to an identification problem. Because every presentation comes from different people, the biggest differences from training sets is quite literally the people involved.

In summary, there are 18 total evaluations conditions. The same 6 training and testing evaluations are used from experiment 2, now across the three conditions: full-contrastive, partially-contrastive and not-contrastive.

7.2.5.1 Exp 3: Results

The deterministic algorithm results are given in Table 7.4 (see next page). The best performing training method, fully contrastive with clean and noise-augmented data, is again indicated by the †. The key finding is noise-augmentations enable partially-contrastive training data to perform in-line with fully-contrastive data. The optimal training method by collection effort is indicated by the ‡. The ‡ approach significantly outperforms the baseline clean fully-contrastive training approach on noisy test imagery and is only slightly worse on clean test imagery. This is a potential huge opportunity to simplify collection costs and timelines. Note a similar phenomenon is seen with not-contrastive data, but it is likely too risky to justify that level of generalization risk.

The deep-learning algorithm results are shown in Table 7.5 (see next page). They follow the same trend, where including noise data in the training dataset can notably improve performance. The best performance again is achieved using fully-contrastive data with noise-augmentations (indicated by the †), and the noise-augmented partially-contrastive dataset again optimizes collection cost (indicated by the ‡). One interesting observation is these results in general show more sensitivity to training data than the deterministic approach. The MobileNetV2 performs better than LBP when given sufficient training data, but also suffers more with not-contrastive data. It is theorized this is because the deep-learning algorithm needs to learn the feature-space from the liveliness imagery (e.g., transfer-learning from ImageNet does not contain spoofing data) whereas the deterministic algorithm only needs to isolate the LBP components associated with liveliness.

Algorithm	Train	Test	ACER	NPCER	APCER
No Participant-Liveliness Contrast (Degraded Dataset):					
LBP	Clean	Clean	3.3%	5.0%	1.7%
LBP	Clean	Noise	32.7%	48.9%	16.6%
LBP	Noise	Clean	3.4%	3.5%	1.7%
LBP	Noise	Noise	16.9%	30.3%	3.4%
LBP	Clean & Noise	Clean	3.3%	4.7%	1.9%
LBP	Clean & Noise	Noise	17.3%	31.6%	3.0%
Partial Participant-Liveliness Contrast (Degraded Dataset):					
LBP	Clean	Clean	2.8%	4.9%	0.8%
LBP	Clean	Noise	33.4%	48.4%	18.3%
LBP	Noise	Clean	2.4%	3.5%	1.3%
LBP	Noise	Noise	5.2%	8.7%	1.7%
LBP ‡	Clean & Noise	Clean	2.2%	3.4%	1.1%
LBP ‡	Clean & Noise	Noise	5.2%	8.8%	1.6%
Full Participant-Liveliness Contrast (Original Dataset):					
LBP	Clean	Clean	2.6%	4.0%	1.2%
LBP	Clean	Noise	33.1%	44.0%	27.9%
LBP	Noise	Clean	2.8%	4.7%	1.0%
LBP	Noise	Noise	4.5%	7.6%	1.4%
LBP †	Clean & Noise	Clean	2.8%	4.6%	1.0%
LBP †	Clean & Noise	Noise	4.5%	7.6%	1.4%

Table 7.4: Evaluating near-infrared material-spectroscopy robustness to dataset contrast-degradation when noise-augmented: deterministic results. Algorithms are trained using 10% camera and environmental noise-augmentations (randomly distributed), then evaluated on clean and noisy data (note noisy-test is 100% noise-augmented). Noise here implies both camera and environmental noise generators. The best performing training method is indicated by the † and the optimal training method by collection effort is indicated by the ‡.

Algorithm	Train	Test	ACER	NPCER	APCER
No Participant-Liveliness Contrast (Degraded Dataset):					
MobileNetV2	Clean	Clean	13.2%	20.8%	5.7%
MobileNetV2	Clean	Noise	22.2%	20.0%	24.4%
MobileNetV2	Noise	Clean	13.6%	15.6%	11.5%
MobileNetV2	Noise	Noise	6.4%	9.5%	3.4%
MobileNetV2	Clean & Noise	Clean	3.6%	4.0%	3.2%
MobileNetV2	Clean & Noise	Noise	4.5%	2.8%	6.1%
Partial Participant-Liveliness Contrast (Degraded Dataset):					
MobileNetV2	Clean	Clean	2.6%	1.2%	4.0%
MobileNetV2	Clean	Noise	13.7%	11.5%	16.0%
MobileNetV2	Noise	Clean	5.6%	6.3%	4.9%
MobileNetV2	Noise	Noise	5.1%	0.2%	10.0%
MobileNetV2 ‡	Clean & Noise	Clean	1.5%	0.2%	2.7%
MobileNetV2 ‡	Clean & Noise	Noise	5.6%	6.3%	4.9%
Full Participant-Liveliness Contrast (Original Dataset):					
MobileNetV2	Clean	Clean	0.6%	0.3%	0.2%
MobileNetV2	Clean	Noise	18.2%	17.9%	18.5%
MobileNetV2	Noise	Clean	0.7%	0.4%	0.9%
MobileNetV2	Noise	Noise	1.7%	0.5%	3.0%
MobileNetV2 †	Clean & Noise	Clean	0.7%	0.7%	0.7%
MobileNetV2 †	Clean & Noise	Noise	1.0%	0.6%	1.4%

Table 7.5: Evaluating near-infrared material-spectroscopy robustness to dataset contrast-degradation when noise-augmented: deep-learning results. Algorithms are trained using 10% camera and environmental noise-augmentations (randomly distributed), then evaluated on clean and noisy data (note noisy-test is 100% noise-augmented). Noise here implies both camera and environmental noise generators. The best performing training method is indicated by the † and the optimal training method by collection effort is indicated by the ‡.

7.3 Conclusion

In conclusion, near-infrared material-spectroscopy algorithms can be sensitive to camera and environmental noise. These noise factors are commonly seen in facial-liveliness-verification, but it is expensive to repeat the data-collection under noisy conditions. Hence, a noise-augmentation paradigm is presented to generate semi-realistic camera and environmental noises. These are physics-informed but not photo-realistic for pragmatic reasons.

Unfortunately, the first experiment shows that the algorithms are by default sensitive to both noises. The algorithms seem more sensitive to environmental noise and this seems reasonable; the camera noises are a slight perturbation to the image, whereas the environment can occlude facial-features. Fortunately, including noise-augmentations in training addresses this concern. The second experiment improves noise robustness without any degradation to clean data accuracy.

The most interesting finding is that training on only noise-augmented data seems to be sufficient. This suggests that using noise-augmentations can potentially reduce the need for real-world data. To evaluate this, a third experiment is conducted on degrading the dataset liveliness-contrast. Two new degradation states are introduced, partially-contrastive and not-contrastive, which intentionally remove liveliness pairs (e.g., live versus spoof). The inclusion of noise-augmentations surprisingly seems to generally enable algorithms trained on partially-contrastive data to keep up with algorithms trained on (clean) fully-contrastive data.

This last finding is potentially very significant. Collecting fully-contrastive data is time consuming and expensive. To only require a small subset of fully contrastive data (here one-third is conveniently selected) by synthetically introducing noise is a win for future research. These benefits are also believed to be conservatively representative, as only 10% of the training data is randomly noise-augmented. This value is selected to keep approximately a 1-1 ratio of clean and noisy images. That said, the benefits here imply going on a more aggressive ratio can potentially further improve performance. This is not evaluated because of computational constraints (execution takes on the order of days per scenario) but hypothesized to be valid.

In conclusion, the recommendation is to always include camera and environmental noise-augmentations when training. This improves robustness with no risk to clean-performance, and can potentially be used to reduce real-world data collection. The only remaining question is whether the algorithms can become completely robust to noise (i.e., zero difference in clean versus noisy performance). While this demonstrated performance is reasonable, there is still opportunity for improvement.

CHAPTER 8

Improving Liveliness Robustness via Auxiliary-Noise-Tasks

The near-infrared (NIR) material-spectroscopy (MS) methodology is thus far shown to be robust across facial-liveliness-verification (FLV) scenarios. First, common user presentation use-cases are evaluated in Chapter-6. This include variations of head-pose, position and ambient-lighting. Next, image noises are introduced through synthetic camera and environment generators in Chapter-7. This evaluation shows the algorithms are sensitive at first, but can improve when including noise-augmentations in the training process. This training exploration actually demonstrates that noise-augmentation can potentially replace real collected-data. This is an exciting finding, which intuitively implies noise-context can improve FLV robustness.

This chapter builds off the intuition and proposes explicitly learning noise-features. The auxiliary-noise-task (ANT) framework is a novel, noise-based multi-task-learning (MTL) approach. The idea is to employ auxiliary-tasks to describe intra-class variance. In essence, variance is problematic because of the potential overlap in liveliness classes. For example, poor exposure can make live and spoof faces appear equally bright. The intention is to utilize ANTs to describe the relevant noises and better describe liveliness. This methodology is specifically applied on the noise-augmentation generators.

The hypothesis is this methodology generates orthogonal features for liveliness and image-noise. This enables separating them within the convolutional-encoder, such that the tasks can isolate only the relevant features. As such, learning noise features ironically functions as a de-noising filter. Jointly learning orthogonal tasks, however, is non-trivial. To ensure parameter convergence, the chapter first experimentally evaluates a series of training practices. Once the optimal training method is identified, the ANT networks are evaluated on the noise-augmented spectroscopy dataset. The goal is to become completely robust to noise; i.e., have no discernible performance difference between clean and noisy test data. Framework utility is then further validated by applying it onto a new dataset that naturally has more image-noise.

8.1 Optimizing Features via Auxiliary-Noise-Tasks

The ANT framework is designed to teach DL networks orthogonal signal and noise features. By describing noises that cause intra-class variance with ANTs, the liveliness features can be isolated. This approach essentially builds a de-noising filter within the convolutional-neural-network (CNN) features, as has the advantage of not needing another algorithm. That is to say the CNN itself does the de-noising rather than requiring pre-processing steps. This computationally efficient and avoids the risk of bias from secondary algorithms.

8.1.1 ANT Network Topology

This research proposes a novel application of multi-task-learning (MTL) to explicitly learn noise. Each labelled noise factor is assigned a corresponding ANT on the liveliness network. These are then jointly trained to encode liveliness and noise features in the convolutional-neural-network (CNN). In theory, the classification task layers can then isolate the features directly associated.

A generic ANT network topology is illustrated in Fig. 8.1. This example is liveliness classification network, where there is a CNN encoder (blue and grey) appended with the FLV task (solid green) and relevant noise ANTs (transparent green). Note how the classification tasks are independent of each other here. This approach is called soft-knowledge sharing.

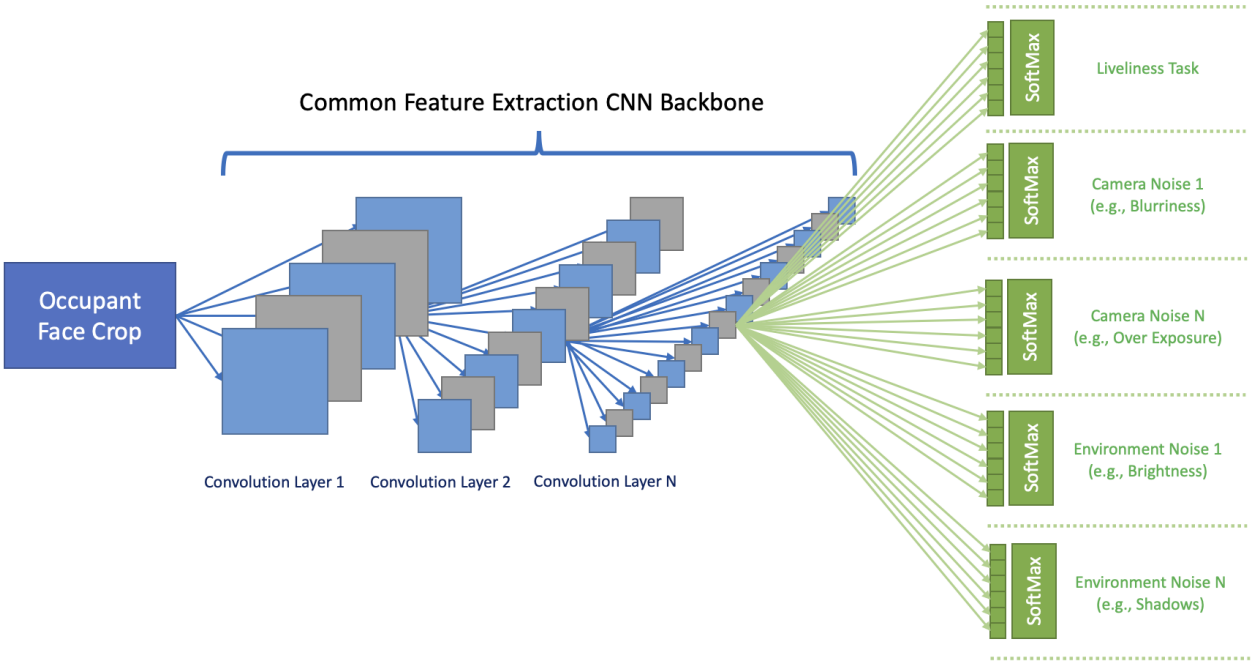


Figure 8.1: Auxiliary-Noise-Task network with soft knowledge-sharing. This network has a primary liveliness task and supporting noise-tasks for camera and environmental noise. Note how there is no interaction with the ANT layers.

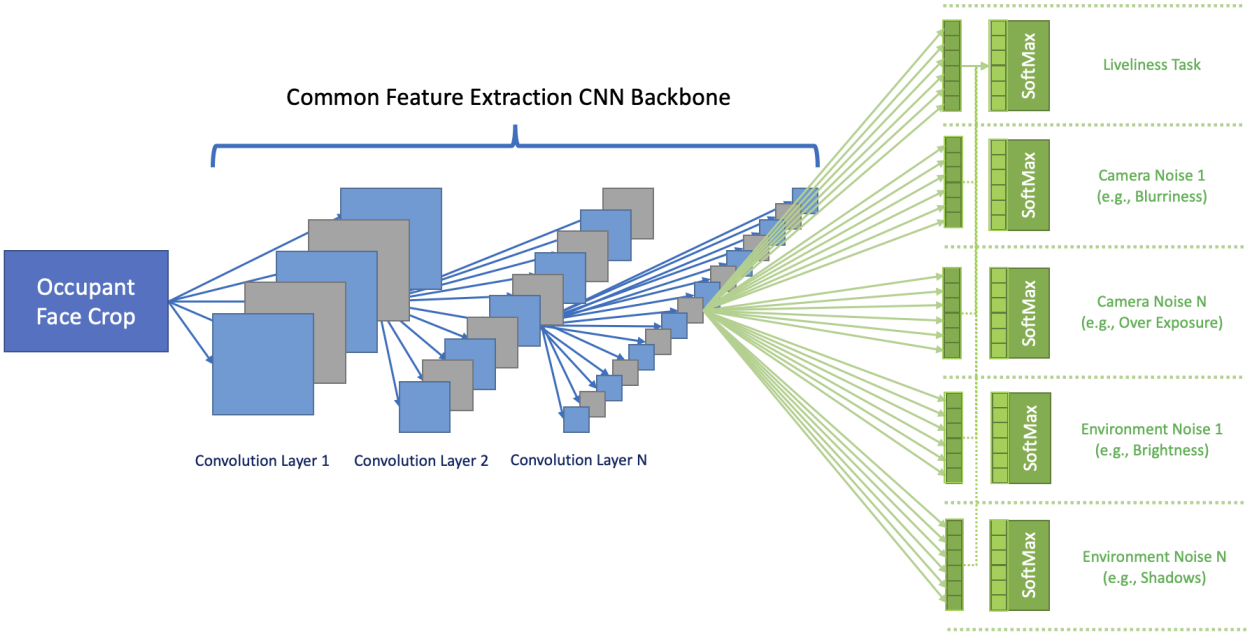


Figure 8.2: Auxiliary-Noise-Task network with hard knowledge-sharing. This network similarly has a primary liveliness task and supporting noise-tasks for camera and environmental noise. Note how there is now a direct knowledge sharing from the ANT layers.

Soft-knowledge sharing has run-time benefits. The tasks can be used for training, then removed at run-time. Note that if some ANTs are useful for the FR, they can be kept in lieu of adding secondary algorithms (e.g., using the liveliness network to estimate image quality).

In some cases, the classification tasks may be constructive, such that direct interaction can improve performance. This approach is known as hard-knowledge-sharing. An example of this approach is illustrated in Fig. 8.2 (see next page). Note how the ANT classification outputs are directly used as an input to the primary task. In this scenario the ANTs must be inferred at run-time regardless of application utility.

8.1.2 ANT Joint Loss

Training knowledge is back-propagated into Neural-Networks through loss functions. Each time a training sample is inferred, the "distance" between the inferred value and training label is calculated and used to adjust the weights. Commonly used loss functions are given in (8.1) and (8.2).

$$MSE = \frac{1}{N_{MSE}} \sum_{i=1}^{N_{MSE}} (y_i - \hat{y}_i)^2 \tag{8.1}$$

$$CE = -\frac{1}{M_{CE}} \sum_{i=1}^{M_{CE}} y_i \log p(y_i) \tag{8.2}$$

In theory any number or classification type of ANT can be used. As an example, a regression ANT would be associated with mean-squared-error loss as applied in (8.1). Here the loss is the distance between inference y_i and label $\hat{y}_i$ for N_{MSE} values. Alternatively, a categorical ANT would employ a log-loss (i.e., cross-entropy) as applied in eq. 8.2. This better separates out the relative distance between classes by incorporating classification probability, p_{CE}, for M_{CE} classes. The only real limitation is the classification difficulty; combining multiple challenging tasks may require more sophisticated loss methods or may not converge.

$$L_{ANT} = \sum_{t=1}^{K_{ANT}} w_t \times L_t(Y_t, \hat{Y}_t) + \eta_t \tag{8.3}$$

The ANT framework combines these losses in joint expected value. The loss combination is applied in (8.3) by summing the losses for each task, $L_t(Y_t, \hat{Y}_t)$ (where Y_t and $\hat{Y}_t$ are the task inference and label vectors). This expected value includes weight w_t and normalization η_t factors for each of the K_{ANT} tasks. It is recommended to associate the loss weight by task difficulty and importance, where normalizing to the number of classes may help. For example, binary-cross-entropy will have loss on a 0 to 1 scale, whereas Mean-Squared-Error is proportional to the number of values square; these should both be normalized to the 0 to 1 range. Improper weighting and normalization can otherwise lead to stopping early or even destructive interference when training.

8.1.3 Specialized Training Scheduling

The ANT framework is designed to learn orthogonal signal and noise features. This runs the risk of destructively interfering with each other, degrading parameter optimization. In theory one can simply bias the loss weights by task priority, but this runs the risk of the other tasks not actually converging (defeating the purpose). For these reasons, a traditional static loss relationship may not work.

A dynamic loss re-balancing strategy is proposed to avoid this. The loss priority should be first biased towards the most difficult task. Once that has stable features (e.g., 95% validation accuracy), the loss weighting is re-balanced to favor the worst performing task. This process of re-balancing is to continue until all tasks demonstrate robust validation accuracy.

This dynamic loss re-balancing should encode the orthogonal features. One risk, however, is the classification layers may not be sufficiently optimized as a result of the joint learning. A fine-tuning process can be applied to address this. Once joint loss is stable, the encoder weights are frozen. Each of the tasks are then independently fine-tuned on the dataset (i.e., other tasks are frozen and have zero loss contribution).

If the classification task is too difficult to converge, a one-class learning approach can be used by introducing noises in iterative training runs. That is to say first the clean data should be learned until validation accuracy is precise, then the training dataset should be appended with the noisy data. All ANT losses should be set to zero when training on clean data to avoid irrelevant bias.

8.2 Performance Evaluation

The ANT framework is designed to isolate liveliness features from image-noise. Three experiments are designed to evaluate the utility of ANT networks for FLV. First, a sensitivity analysis is conducted to identify the optimal ANT network topologies and training method. Once the optimal method is identified, the ANT network is evaluated on the noise-augmented spectroscopy dataset. Lastly, the utility of the framework is further validated by applying it onto another dataset.

8.2.1 Validation Algorithms

All ANT networks are evaluated without the contextual tasks. Given prior results show the deep-learning algorithms are already more accurate than the best evaluated deterministic algorithm local-binary-pattern (LBP), no further validation is viewed as necessary.

8.2.1.1 Network Without Joint-Learning

The first validation approach is to use the state-of-the-art liveliness network without the contextual tasks. For FLV, the face-crop is described by a convolutional-neural-network (CNN) feature encoder then classified using a 128-neuron feature fully-connected layer and 2-neuron classification layer. The MobileNetV2 (108) and InceptionV3 (109) encoders are considered for an efficient and robust network respectively. Both encoders are pre-trained on ImageNet (110) and fine-tuned on the evaluation dataset. Training and evaluation use the Tensorflow toolbox (111).

8.2.2 Research Limitations

Multi-task-learning is a new branch of deep-learning. At the time of this research, there is minimal art on face-recognition applications and none identified for face-liveliness. From this perspective this research does not have a true validation benchmark except to compare the network topologies. Furthermore, the attack presentations and noises selected are commonly found but inherently not fully inclusive. A general pragmatic approach to algorithm development would be to verify it meets requirements on this evaluation dataset, then make updates based upon penetration testing.

Noise	Generator
Camera focus[1] (blurriness)	Gaussian Low Pass Filter
Dark noise[1] (random leakage)	Gaussian Noise Generator
Shot noise[1] (random photo distribution)	Poisson Noise Generator
Salt and pepper noise[1] (analog-to-digital error)	Random 0 and 255 Generator
Under-exposure[1] (low contrast)	Gamma Subtraction
Over-exposure[1] (saturation)	Gamma Addition
Point-source[2] (point sources)	Synthetic Bright Ellipse
Point-shadow[2] (shadows)	Synthetic Dark Ellipse
Streaking-source[2]	Synthetic Overhead Sun
Streaking-shadow[2]	Synthetic Overhead Shadow
Piping-source[2]	Synthetic Side Sun
Piping-shadow[2]	Synthetic Side Shadow

Table 8.1: Auxiliary-noise-tasks: camera and environmental noises. This table enumerates both the relevant types of noises and how they are being generated for evaluation. Camera noises are indicated by [1]. Environmental noises are indicated by [2].

8.2.3 Exp 1: Auxiliary-Noise-Task Topology and Training Optimization

The first experiment is designed to identify the optimal ANT network topology and training methods. Getting orthogonal signal and noise tasks to converge is non-trivial (no relevant literature is identified). Hence, the network topologies and training-loss methods are evaluated, along with the training-loss strategies. This is done on the not-contrastive, noise-augmented dataset from Chapter-7.2 to best see the performance differences.

The camera and environmental noise-augmentations are recapped in Table 8.1. Recall these are physics-informed but semi-realistic. Noise-augmentations are applied during (randomly selected) to maintain approximately a 1:1 ratio of clean and noise-augmented training data. All test-images are fully noise-augmented for completeness. The network optimization is then evaluated by varying the topologies and training strategies. These variables are given in Table 8.2.

Encoders	Auxiliary Noise Tasks	MTL Sharing	Training Loss
MobileNetV2	Cam Noises	Soft Sharing	Static
	Env Noises	Hard Sharing	Dynamic
	Cam & Env Noises		Dynamic + Tuning

Table 8.2: Auxiliary-noise-task network topology and training variables: near-infrared material-spectroscopy application. A network is generated for each topology and training-loss.

For simplicity, the only the optimal MobileNetv2 encoder is used. There are 18 distinct network topology and training-loss combinations. These are evaluated 10 times each using stratified cross-validation for a total of 180 networks. Performance is reported for each structure as an average of cross-validation test accuracy. These ANT networks are benchmarked against the base networks without ANT and top deterministic NIR MS algorithm (local-binary-patterns). Evaluation metrics are nominal-presentation-classification-error-rate (NPCER), attack-presentation-classification-error-rate (APCER) and average-classification-error-rate (ACER). These are essentially the false-rejection-rates of live people, false-acceptance-rates of spoofs and the average of both. Note that ACER is utilized over test-average as the dataset is imbalanced.

8.2.3.1 Exp 1: Results

The topology and training analysis shows the inclusion of ANTs significantly improves performance. The benefits of including the camera and environmental ANTs are visualized in Table 8.3. Recall that this experiment uses the non-contrastive dataset (worst conditions). Note how all of the ANT networks notably outperform the base network.

Encoder	MTL	Loss	Test	ACER	NPCER	APCER
Base Network (No Auxiliary-Noise-Tasks):						
MobileNetV2	None	Static	Clean	13.2%	20.8%	5.7%
MobileNetV2	None	Static	Noisy	22.2%	20.0%	24.4%
Camera and Environmental Auxiliary-Noise-Task Networks:						
MobileNetV2	Soft	Static	Clean	2.9%	5.1%	0.6%
MobileNetV2	Soft	Static	Noisy	6.0%	5.5%	6.5%
MobileNetV2	Soft	Dyn	Clean	1.1%	0.2%	2.1%
MobileNetV2	Soft	Dyn	Noisy	2.4%	3.1%	1.7%
MobileNetV2 †	Soft	Dyn + Fine	Clean	0.5%	0.8%	0.2%
MobileNetV2 †	Soft	Dyn + Fine	Noisy	0.9%	1.5%	0.2%
MobileNetV2	Hard	Static	Clean	3.4%	2.4%	4.5%
MobileNetV2	Hard	Static	Noisy	2.3%	3.5%	1.1%
MobileNetV2	Hard	Dyn	Clean	1.1%	1.9%	0.2%
MobileNetV2	Hard	Dyn	Noisy	1.9%	1.6%	2.2%
MobileNetV2	Hard	Dyn + Fine	Clean	4.2%	0.1%	8.2%
MobileNetV2	Hard	Dyn + Fine	Noisy	3.5%	4.1%	2.9%

Table 8.3: Evaluating optimal auxiliary-noise-task network topology: synthetic camera and environmental tasks. It is clear the ANT framework improves performance, where the best combination is indicated by the †. Dynamic scheduling is abbreviated to Dyn. Fine tuning is abbreviated to Fine.

Recall from the NIST ISO 30107 that an APCER of 5% is typically recommended; almost all of the ANT networks achieve this. The optimal combination seems to be soft-sharing with the dynamic loss-weighting and fine-tuning (indicated by the †). This intuitively makes sense. The training schedule is most important as these tasks are designed to be orthogonal, meaning the losses for the associated tasks need to be dynamically re-balanced for optimal encoding. This practice, however, can result in unrefined fully connected layers. The final fine-tuning step addresses this issue. Furthermore, the tasks being orthogonal means hard knowledge sharing should provide minimal benefits. While hard knowledge-sharing can theoretically improve performance, it is often difficult to implement well because the task-losses are inherently muddled. It is difficult to de-couple errors from the primary task features versus the knowledge shared from the secondary tasks. Given there is minimal benefits to sharing, it results in generally being destructive. Note that for space reasons only the networks utilizing both camera and environmental ANTs are shown here; the camera versus environment ANT evaluation is presented in Appendix-D (where camera tasks are generally more useful but combining both is best).

8.2.4 Exp 2: Applying Auxiliary-Noise-Tasks to Address Contrast

The results from the first experiment make it clear there are benefits to utilizing the ANT framework. The second experiment now applies the optimal combination to address degraded-contrast datasets. If synthetic noise-augmentations with ANT can enable partially-contrastive data to meet intended robustness, researchers can dramatically save on real-world data collection efforts.

This experiment essentially a repeat of the dataset-contrast sensitivity study. The participant data is intentionally degraded. Two new contrast-degraded states are introduced: partially-contrastive and not-contrastive. The partially-contrastive dataset randomly selects between live, spoof or live and spoof presentations for training. That is to say all the data exists, but only a subset of liveliness classes are selected to ensure roughly one-third are fully contrastive. The not-contrastive dataset randomly selects between live or spoof presentations for training - but no participants have all presentations.

The same scenarios are now evaluated with the optimal ANT framework. Training is varied between clean data only, noise-augmented data only and clean with noise-augmented data. Testing is conducted on clean and noise-augmented data independently. This results in 6 scenarios evaluated on fully-contrastive, partially-contrastive and not-contrastive datasets. These are evaluated 10 times using stratified-cross-validation for totality of 180 networks.

8.2.4.1 Exp 2: Results

The evaluation results of the ANT framework on contrast-degraded datasets are given in Table-8.4. These are outstanding. Regardless of degradation state, the MobileNetV2-ANT algorithm achieves excellent results on clean and noisy data.

Observe how the ANT network trained on partially-contrastive (indicated by the †) data actually outperforms the base-network trained on fully-contrastive data. This is a substantial contribution: utilizing noise-features can improve performance while saving on data-collection costs. If outright performance is the primary objective, the strongest algorithm indicated by the ‡. That network essentially makes zero mistakes in test inference.

Algorithm	**Train**	**Test**	**ACER**	**NPCER**	**APCER**
No Participant-Liveliness Contrast (Degraded Dataset):					
MobileNetV2	Clean	Clean	13.2%	20.8%	5.7%
MobileNetV2	Clean	Noise	16.2%	20.6%	12.2%
MobileNetV2-ANT	Noise	Noise	2.3%	0.8%	3.8%
MobileNetV2-ANT	Noise	Clean	1.1%	0.7%	1.4%
MobileNetV2-ANT	Clean & Noise	Clean	0.3%	0.5%	0.1%
MobileNetV2-ANT	Clean & Noise	Noise	1.0%	1.2%	0.9%
Partial Participant-Liveliness Contrast (Degraded Dataset):					
MobileNetV2	Clean	Clean	2.6%	1.2%	4.0%
MobileNetV2	Clean	Noise	16.8.%	21.7%	11.8%
MobileNetV2-ANT	Noise	Noise	1.4%	1.0%	1.8%
MobileNetV2-ANT	Noise	Clean	1.1%	0.6%	1.6%
MobileNetV2-ANT †	Clean & Noise	Clean	0.2%	0.4%	0.1%
MobileNetV2-ANT †	Clean & Noise	Noise	0.8%	1.4%	0.2%
Full Participant-Liveliness Contrast (Original Dataset):					
MobileNetV2	Clean	Clean	0.6%	0.3%	0.2%
MobileNetV2	Clean	Noise	14.5%	17.4%	11.6%
MobileNetV2-ANT	Noise	Clean	0.2%	0.2%	0.1%
MobileNetV2-ANT	Noise	Noise	0.5%	0.3%	0.7%
MobileNetV2-ANT ‡	Clean & Noise	Clean	0.2%	0.1%	0.3%
MobileNetV2-ANT ‡	Clean & Noise	Noise	0.1%	0.0%	0.2%

Table 8.4: Addressing dataset contrast via the auxiliary-noise-task framework. The optimal ANT algorithm, MobileNetV2 trained with camera and environmental ANTs using dynamic scheduling with fine tuning, is evaluated on the degraded-contrast datasets. The optimal algorithm by data-collection needs is indicated by the †. The best performing algorithm is indicated by the ‡.

8.2.5 Exp 3: Applying to Occupant-Monitoring Perspective

To further evaluate the utility of ANTs, a new dataset is acquired from the research sponsor, Ford Motor Company. This dataset is designed for cabin monitoring. The camera is offset to the middle of the vehicle (versus being directly head-on) and introduces the RGB-IR color-filter-array (CFA). This CFA is able to image color and near-infrared (NIR) photons. In principle, this may seem like an advantage. One can imagine that utilizing features from the visible spectrum could be beneficial (especially for face-identification). In practice, this adds visible noise to the NIR pixels due to the filter design. This noise is potentially problematic for spectroscopy. Furthermore, NIR pixels are now only one-quarter of the full resolution, meaning there is inherently less information for texture classifiers to work with.

An example from the Ford dataset is shown in Fig. 8.3. This dataset is comprised of 58 participants are imaged while driving a vehicle. The intention here is emphasized towards identification noise factors; as such, participants are coached to vary their head-pose and utilize accessories (e.g., hat, wigs and cell phone). Furthermore, the drives are structured to vary sun angle such that there are various bright spots and shadows introduced. In totality this dataset has approximately 50,000 unique frames. Only the driver's face is considered from each frame (cropped with Retina Face (12) and resized to algorithm inputs).

Figure 8.3: Occupant-monitoring RGB-IR perspective sample. This camera technology is more sensitive to ambient light and generally harder to see the illumination.

Encoders	Auxiliary Noise Tasks	Knowledge	Training Loss
MobileNetV2	Identity	Soft Sharing	Static
InceptionV3	Covid-mask	Hard Sharing	Dynamic

Table 8.5: Auxiliary-noise-task network topology variables: occupant-monitoring RGB-IR application. A network is generated for each topology and training-loss.

Similar to the first topology evaluation, it is important to identify the correct ANT combinations. This dataset is already noisy (by material-spectroscopy standards), but there are no explicit labels for image effects. Furthermore, the dataset is not balanced. There are significantly more live presentations, and not all participants have a matching spoof (in this case paper-mask).

An alternative set of ANTs are proposed for this dataset. The identity of the person is a liveliness noise-factor when there are not enough contrastive pairs. Furthermore, a very small number of participants are wearing a medical covid-mask. This is not a spoofing mask and therefore needs to be considered a live presentation. These tasks are evaluated in a similar topology assessment, described in Table 8.5. This yields 16 network topologies. These 16 topologies are evaluated 10 times each using stratified-cross-validation for a total of 160 networks. The same performance metrics are utilized here. Note this fine-tuning is no longer included as an option due to needing images of the same participant in the validation set for the identity task. This is presents risk of over-fitting (which is observed in a few anecdotal evaluations).

8.2.5.1 Exp 3: Results

The ANT framework yet again delivers performance benefits. The MobileNetV2 results are given in Table 8.6 (see next page). The best performing efficient network (indicated by the †) incorporates both identification and covid-mask tasks. Like the first sensitivity experiment, the best networks use soft knowledge-sharing and dynamic loss scheduling. This is intuitive for the same reasons as the prior experiment. Loss weighting needs to reflect task convergence, but parameter back-propagation gets complicated when fusing tasks.

An interesting observation is the covid-mask ANT appears to have the most influence. Regardless of knowledge sharing or loss scheduling methodology, this trend generalizes. It was originally theorized that the participant-identification task would be more useful, as the largest variance in the liveliness classes is the person. However, this is can be rationalized by noting there are comparatively few presentations of the covid-masks. Furthermore, it is essentially a hybrid of live and spoof presentations. Regardless, the key finding is that the ANT framework can still deliver performance benefits without noise labels.

Encoders	ANTs	MTL	Loss	ACER	NPCER	APCER
MobileNetV2	None	None	Static	5.0%	5.0%	4.9%
MobileNetV2	ID	Soft	Static	3.5%	1.3%	5.7%
MobileNetV2	Covid	Soft	Static	3.0%	1.1%	4.9%
MobileNetV2	ID & Covid	Soft	Static	5.0%	2.7%	8.3%
MobileNetV2	ID	Soft	Dynamic	4.5%	2.0%	7.0%
MobileNetV2	Covid	Soft	Dynamic	2.9%	0.3%	5.6%
MobileNetV2 †	ID & Covid	Soft	Dynamic	2.6%	0.9%	4.3%
MobileNetV2	ID	Hard	Static	5.0%	2.6%	7.5%
MobileNetV2	Covid	Hard	Static	4.0%	2.3%	5.7%
MobileNetV2	ID & Covid	Hard	Static	4.0%	1.7%	6.3%
MobileNetV2	ID	Hard	Dynamic	5.1%	1.4%	8.8%
MobileNetV2	Covid	Hard	Dynamic	9.6%	1.1%	18.1%
MobileNetV2	ID & Covid	Hard	Dynamic	3.1%	%	4.7%

Table 8.6: Auxiliary-noise-task framework liveliness results: efficient occupant-monitoring network. The best performing ANT topology is given by the †.

For space reasons, the robust network results are given in Appendix-D. It is observed the general trends hold true with the InceptionV3 encoder, but the benefits are diminished. This is theorized to be a result of the InceptionV3 encoder generating a sufficient feature space without the ANTs, therefore adding the secondary tasks only adds complication to the liveliness convergence. This is particularly true when factoring the lack of fine-tuning (due to utilizing identification tasks). It is hypothesized that having a precise noise-labels would yield better gains.

8.3 Conclusions

This chapter demonstrates that encoding noise-features can dramatically improve facial-liveliness-verification performance. Classifiers historically learn to discern features from noise by employing large, contrastive datasets. The auxiliary-noise-task (ANT) framework is designed to instead provide contrast through better labelling. Specifically, the concern is texture-noises can degrade spectroscopy algorithm performance. Hence, the goal is to jointly learn liveliness and noise features, which should be orthogonal in nature. The ANT framework utilizes this orthogonality to ironically de-noise the classification tasks.

Jointly learning orthogonal tasks is not trivial. The network topology and training methods evaluation reveals that tasks should not interact (soft knowledge-sharing) and a dynamic loss-weight

re-balancing strategy with fine-tuning is necessary. These observations are intuitive. If the tasks are truly orthogonal, they need to be learned in a fashion that minimizes interaction but also reflects information from both signal and noise. This said, the ANT networks dramatically outperform the base networks regardless of topology or training approach. These results extend to the point where the best ANT network essentially has zero errors when trained on fully-contrastive data. This phenomenon is also observed on a new dataset provided by the research sponsor, validating the framework's value.

This value is best demonstrated when considering contrast-degraded datasets. The ANT liveliness network trained on partially-contrastive data actually outperforms the baseline network trained on fully-contrastive data. This is a substantial accomplishment, improving performance while also decreasing data-collection needs. This is a rare feat and makes the ANT framework a success.

This work concludes the physical-spoof-attack investigations. The proposed method is robust to every considered noise-factor and does so with data-collection optimizations. All future work proposals are discussed in the conclusion (Chapter-11).

CHAPTER 9

Image-Integrity-Verification via Camera-Noise

Photo-realistic tampering is the other key face-recognition (FR) vulnerability. These attacks are designed to modify the image-stream through an injection device, bypassing traditional data security measures (e.g., message authentication). Analogous to the physical-spoof-attack, the face-swap-attack (FSA) replays an acquired headshot (typically from social-media) for authentication. In this scenario the replay is done using face-swapping algorithms. These swaps can be highly-realistic, fooling state-of-the-art identification algorithms (26). Alternatively, attackers can also perform the service-denial-attack (SDA) by discretely removing faces from the imagery. This prevents authentication, either as a means of causing inconvenience or a tool for blackmail. For more details on these attacks, see the threat-model in Chapter-3.

The challenge with photo-realistic tampering is the obvious: the alterations can be imperceptible. These are by design more sophisticated than traditional cut-and-splice methods; as such, new detection methods are required. One approach is to identify traces from the generative-adversarial-networks. These can be effective, though have a tendency to not generalize well (as each generator has unique traces) (83; 80). Other approaches identify temporal anomalies, as it is very difficult to properly represent natural human behavior (45; 84; 85). These approaches also are insufficient, as they are computationally expensive and require the live person to present realistic movements (conflicting with the seamless expectation).

This chapter proposes an efficient "noiseprint"-verification to mitigate the FSA and SDA. Whether introducing or removing faces, the image noise-profile is necessarily altered with tampering. More specifically, noises that can be used as source-identification fingerprints are known as "noiseprints." This proposal securely enrolls the FR platform's camera "noiseprint" for integrity-verification; this is possible because most platforms use an embedded camera. Each time a new frame is presented, the "noiseprint" is estimated then verified against the enrollment for authenticity. This methodology is extremely precise and uses compression to operate with imperceivable latency. Benchmarking is done with three state-of-the-art algorithms to demonstrate a novel combination of accuracy and speed.

9.1 Compressed Photo-Response-Non-Uniformity Analysis

The proposed "noiseprint" for integrity-verification is photo-response-non-uniformity (PRNU). PRNU, an estimation of camera photo-receiver imperfections, is known to uniquely characterize cameras in large scale source identification (73). Given this feature uniqueness, it is intuitive that image modifications should necessarily cause an observable deviation. This intuition is implemented using a "noiseprint" verification process. The FR camera is first securely enrolled, where future frames are then verified for PRNU integrity. Compression is also introduced in the form of down-sampling. This improves run-time though can remove some relevant features. To retain sensitivity, the verification is done over a series of sub-zones.

9.1.1 Photo Response Non-Uniformity Estimation

The PRNU calculation employs the methodology presented by Goljan et al. (73). In general, an image can be described as the sum of the incident light received by the camera, artifacts introduced by camera intrinsics, and temporal noise. This is expressed in (9.1), where the incident light is represented by I_0, the camera noise (PRNU) is represented by K and other noises (quantization, shot, dark current, temporal, etc.) are represented by θ:

$$I = I_0 + I_0 K + \Theta \tag{9.1}$$

To isolate the "noiseprint" the image must be first filtered to remove noise θ. This can be done by applying a Wiener filter, F, to generate residuals $W^i = I^i - F(I^i)$ (75). Note that in this case dark current is assumed to be negligible due to having sufficient scene signal. From there K can be isolated by applying a maximum likelihood estimator (75):

$$\hat{K} = \frac{\sum_{i=1}^{N} W^i I^i}{\sum_{i=1}^{N} (I^i)^2} \tag{9.2}$$

9.1.2 Peak Correlation Energy

One way to classify the camera noise source is to utilize peak-correlation-energy (PCE). The PCE value is computed using the Matlab code provided by Goljan et al. (73). This approach calculates the Pearson's correlation coefficient and then identifies the maximal value for a given sliding window (73). This is achieved by first computing the noise residuals of the hypothesis camera, X, and the image, Y, as a sum of PRNU, K, and secondary noises Θ:

$$\begin{aligned} X &= I\hat{K} \\ Y &= IK_{image} + \Theta_{image} \end{aligned} \tag{9.3}$$

The correlation is computed over shifted areas, where the maximal number of shifts is defined as the product of the difference in image $m \times n$ versus fingerprint dimensions $m_k \times n_k$: $max = (m_k - m + 1)(n_k - n + 1)$ (73) , that is:

$$\rho(s_1, s_2; X, Y) = \frac{\sum_{k=1}^{m} \sum_{k=1}^{n} \big(X[k,l] - \bar{X}\big)\big(Y[k+s_1, l+s_2] - \bar{Y}\big)}{|X - \bar{X}||Y - \bar{Y}|} \tag{9.4}$$

By definition, this implies that PCE is the correlation value for which the peak occurred for shift vector, $s_{peak} = [s_{1max}, s_{2max}]$. In (73), Goljan et al. suggested having the local peak area, η, to be an 11×11 pixel grid. This is given in (9.5):

$$PCE_k(X, Y) = \frac{\left(X \cdot Y(s_{peak})\right)^2}{\frac{1}{mn - |\eta|} \sum_{(s_1, s_2) \in \eta} \left(X \cdot Y(s_1, s_2)\right)^2} \tag{9.5}$$

9.1.3 Source ID: Zonal Expected Value

It is postulated that camera source identification can be improved by dividing the image into sub-zones and taking an expected value. The premise is that the PCE reflects the maximal PRNU similarity in a local window; hence, taking the similarity score over a multitude of zones provides relevant secondary distribution information. This can be thought of as analogous to a Taylor series, where the sub-peaks provide extra harmonics for minimal extra computation.

To calculate ID via zonal-expected-value (ZEV), the hypothesis camera and challenge image PRNU are indexed by zone row and column, X_{rc} and Y_{rc} respectively. A uniform zone distribution is assumed for simplicity. However, in practice, an asymmetric pattern could be leveraged to place emphasis on the most sensitive region of the image. ZEV calculation is then described in Eq. (9.6).

$$ID_{ZEV} = \frac{1}{Z} \times \sum_{r,c \in Z} PCE_K(X_{rc}, Y_{rc}) \tag{9.6}$$

The hypothesis camera that has the highest ID score is then selected as the source.

9.1.4 Tampering-Score: Face-Swap-Verification

Detected faces can be secured through face-zone-verification (FZV). This is done by isolating the face-centroid's zone, z, and applying a tampering-score on it. Tampering-score is generated by applying a tampering filter on the correlation energy. The high pass cutoff is calibrated to ignore standard noise as tampering; the low pass cutoff is calibrated to ignore images that do not match the source. This calibration is done on a per camera basis; in this case a 1% false acceptance rate is prioritized. The FZV score is described in Equation (9.7):

$$T_{FZV} = BP_z(PCE_K(X_z, Y_z)) \tag{9.7}$$

The FZV score offers a significant run-time improvement at the cost of some robustness (as only the facial-pixels are evaluated pixels). In theory, this can be expanded to a nearest neighbor approach, and also to evaluate all zones directly adjacent to the centroid zone z.

$$T_{DZV} = BP_z(K_{rc}), rc \in Z \tag{9.8}$$

9.1.5 Tampering-Score: Service-Denial-Verification

This tampering-score can be used to verify full image-integrity in a ZEV fashion to mitigate face-removal. The tampering filter can be applied on each zone, designated BP_z, where a final filter is applied on the averaged output, BP_I. Each zone filter can be individually calibrated, though for pragmatism uniform cutoffs are assumed. The complete filtered score is described in Equation (9.9):

$$T_{ZEV} = BP_I \left(\frac{1}{Z} \times \sum_{r,c \in Z} BP_z(PCE_K(X_{rc}, Y_{rc})) \right) \tag{9.9}$$

If the image appears tampered in the expected value, it is labelled as tampered. Otherwise, it is labelled authentic.

9.1.6 Compression via Down-Sampling

Compression is introduced in the form of down-sampling to optimize run-time performance. Down-sampling is chosen because it is computationally cheap, and implicitly behaves as an averaging filter when combining neighboring pixels. This will additionally shrink the memory footprint proportionately to the amount of down-sampling. The goal is to make this verification algorithm available for embedded security systems, where resources are often constrained.

All images are evaluated using three compression settings. These are enumerated below, with the axis sampling rate provided in the form of (row, column):

1. Full-Scale Resolution (1x1)
2. Quarter-Scale Resolution (1/2 x 1/2)
3. Sixteenth-Scale Resolution (1/4 x 1/4)

As a point of comparison, two approaches of PRNU enrollment template compression are also evaluated. Pre-compression is defined as acquiring PRNU template on a full-scale image and then

down-sampling. Post-compression is defined as first down-sampling the image and then acquiring the PRNU template. Both approaches are validated in Experiment 1 regarding compressed camera source identification (Section: 9.2.4). Once the optimal compression process is determined it is used for all subsequent experiments.

9.2 Performance Evaluation

The efficacy of the proposed framework is evaluated on both public and private datasets using four experiments. The first two experiments investigate the uniqueness of camera PRNU when applying down-sampling. These experiments are evaluated on a public dataset containing different cameras (Dresden (119)) and a custom dataset composed of cameras of the same makes, respectively.

The second set of experiments investigate detection of image tampering attacks. Tampering is done in the form of the face-swap-attack (FSA) and service-denial-attack (SDA). The FSA is where the attacker digitally swaps in the face of an enrolled user for authentication purposes. This attack is introduced using manual and artifial-intelligence tools. The SDA is where the attacker removes the face from the image, to deny access. This attack is simulated using blob swaps (for general tampering detection sensitivity).

9.2.1 Run-Time Metrics

Run-time is evaluated to pragmatically optimize for real-time applications. Metrics include the average computation time (in seconds) and classifier memory space (in Megabytes). For simplicity, it is assumed that the image is already read into memory where only the PRNU calculation and the camera classification time are measured. These experiments are conducted utilizing a Dell Latitude E5570 laptop; only a single CPU core is employed without any hardware acceleration via co-processors (i.e., no GPU, DSP, etc.).

9.2.2 Open-Source Validation Algorithms

To benchmark the performance of this tampering score, three open-source algorithms are evaluated on the datasets. The algorithms are selected based on both their relevance and their availability on GitHub.

9.2.2.1 Error-Level-Analysis

The first algorithm is a deep learning approach leveraging error-level-analysis (ELA). Gunawan et al developed this approach by transforming images using ELA and feeding a light weight convo-

lutional neural network examples of authentic and spliced images from an open-source tampering dataset (120). For this evaluation the network is retrained on the same experimental data. This algorithm is referred to as "ELA CNN."

9.2.2.2 Image Forgery Detection Tool

The second algorithm comes from the ensemble Image Forgery Detection Tool (79). Levandoski and Lobo present deterministic algorithms for re-compression, color filter array anomalies, noise variance and generic image duplication due to copying (79). It is key to note that due to run-time issues the color filter array analysis is omitted as the algorithm takes over 30 minutes per image on this research machine and is not viewed as real-time. Given there that is no calibration setting, the algorithm suite is run directly on the experimental data without modification. This algorithm is referred to as "Forgery Tool."

9.2.2.3 Discrete-Wavelet-Transform

The third algorithm is the discrete-wavelet-transform (DWT) for blind image tampering (78). While a slightly older paper (2009), Mahdian and Saic apply a methodology that aligns very similarly with this paper's PRNU analysis, the key novelty is the inclusion of apriori camera knowledge. For this evaluation a tampering threshold is calibrated on the same experimental data. This algorithm is referred to as "DWT."

Note that run-time evaluation is not performed on these open-source algorithms because they are written in a different language than the proposed framework. A fair comparison cannot be done across applications.

9.2.3 Research Limitations

It is acknowledged that this verification approach does require apriori knowledge of the camera intrinsics. Real-time facial recognition is typically a closed-loop system, where the camera is integrated into the system and knowledge of the sensor can be easily acquired. This said, even distributed systems can be modeled in this way. The template could be acquired from other representative samples. For example, social media platforms can perform image tampering analysis by using other images uploaded by the user (or their canvas of users).

Furthermore, the lack of open-source tampering datasets that include camera sources requires the comparative algorithms to be validated on this experimental data. While all competitor algorithms are re-trained/re-calibrated to the experimental data, their designs do not necessarily assume blob or face swap detection. From that perspective it is a valid comparison, but it is important to acknowledge that the original authors did not perform their own tuning.

Lastly, it is relevant to note there are real-world noise factors that this study did not introduce. For example, social media applications may employ their own proprietary compression methods to streamline data transmission; Meij et al. have started this investigation specifically for the "whatsap" communication tool (121), noting some degradation is to be expected. Additionally, real world camera exterior camera applications, such as security monitoring, introduce environmental noises (e.g., ambient light, rain, dirt, dust, heat, etc.). Validating the impact of the noises would further justify this methodology, as well as potentially identify other novel control methods.

9.2.4 Exp 1: Compressed Source Identification - Different Cameras

Given that the methodology leverages compression to improve run-time performance, it is important to first verify that PRNU features can be reliably retained when down-sampling. This experiment utilizes a public dataset to evaluate some known worst-case identification scenarios, applying down-sampling and validating the resulting mean average precision.

Worst-case conditions are selected from the results of Dr. Fridrich's large-scale camera identification study (73). They identify that PRNU is least separable when the images analyzed are of the same scene, and if they are taken using the same camera model. To provide a common reference, the Dresden camera forensics image database is utilized (119). While the Dresden dataset does not have cameras of the same make, it does have several families of models. For this reason, the Nikon family of cameras is selected to reflect images of same scene with similar models. 150 images are utilized, with an 80/20 training ratio. This dataset is designated as Dresden-Nikon.

The camera source identification is then evaluated using the down-sampling and compression approaches described in the methodology. The image source is then identified using ZEV classification, for all combinations of zones, down-sampling and compression process (indicated by "Pre-Compression" and "Post-Compression," respectively). The selected classification performance metric is the mean average precision of source identifications.

9.2.4.1 Exp 1: Results

PRNU estimation is in fact sensitive to the compression-order. When first compressing the training image and extracting the PRNU, there is significant degradation in PCE score (resulting in poor mean average precision). However, by acquiring the PRNU first and then compressing the template, much of the information is retained. It is noted that these results also imply that some features are lost when compressing prior to classification for run-time purposes. The lack of classification degradation is explainable by this being a "verification" problem; the challenge camera needs only to be verified that against the expected source. Given that the training template is robust, the peak correlation analysis can tolerate noise added to the challenge.

Compression Factor	**Single Zone** (% Correct)	**16 Zones** (% Correct)	**100 Zones** (% Correct)
Full-Scale	100%	100%	100%
Quarter-Scale (Pre-Compression)	71.7%	73.3%	77.5%
Sixteenth-Scale (Pre-Compression)	38.3%	42.5%	45.8%
Quarter-Scale (Post-Compression)	97.5%	98.3%	99.2%
Sixteenth-Scale (Post-Compression)	97.5%	97.5%	97.5%

Table 9.1: Dresden-Nikon dataset classification performance.

The lossy-compression source-identification accuracy (different cameras) is given in Table 9.1. It can be observed that dividing the image into sub-zones can incrementally improve performance. The performance improvements become more pronounced as the image degrades; this can be seen by juxtaposing the classification rate across image compression. This is then further exhibited when using low-resolution sensors in Experiment 2, where the methodology is applied on Raspberry Pi cameras. Given the simplicity of this approach, these benefits are considered a good return on computational investment.

9.2.5 Exp 2: Compressed Source Identification - Same Cameras

The second experiment expands on these results by addressing the use case of having cameras of the same model. A new database is constructed using 3 identical Ras Pi cameras (for reference, these have a default resolution of 1920×1080). That is to say, they are of the same make and model, and any variation is a result of manufacturing tolerances only. This inherently a harder use-case. Even state-of-the-art source-identification methods indicate that having identical components reduces the "noiseprint" variability (73).

This Ras Pi Camera dataset is collected by imaging 150 photos from each device under four lighting conditions. The same performance evaluation approach is then applied. The dataset is broken into an 80/20 split for training and testing. same 80/20 training ratio is employed. The calculated metric is source-identification mean-average-precision. This is calculated when using ZEV on all down-sampling, resolution and compression-order combinations. It is anticipated that performance will decrease as there is less variance between classes.

9.2.5.1 Exp 2: Results

The compressed source-identification accuracy using the same camera-models is given in Table 9.2. This scenario is clearly harder than when evaluating different cameras-models, as source-identification accuracy is down across the board. One clearly observation is the compression order matters now. Generating robust PRNU features requires first estimating the "noiseprint," then applying the down-sampling. This does not retain all features, but is clearly better than down-sampling first then estimating PRNU. This is likely because down-sampling is effectively low-pass filtering. Given PRNU is a micro-noise, any sort of filtering runs the risk of removing key information. Note that this reasoning is why many state-of-the-art tampering detection networks recommend using patch analysis. Resizing the image to fit the deep-learning inputs inherently runs the risk of feature-removal (81).

The benefits of ZEV are clearly obvious here. When the features are muted, the secondary distribution-correlation helps notably improve classification accuracy. This speaks to the utility of the methodology, adding sensitivity without the risk of over-fitting to random noises. The minimal computation required makes this a very good return-on-investment.

In summary, this experiment confirms two things. First, the optimal training procedure should acquire the full-scale enrollment PRNU and then compress it. Second, the zonal-analysis methodology should be used when local-sensitivity is necessary; this is particularly relevant when considering photo-realistic tampering. These practices are applied in the next experiments.

Compression Factor	**Single Zone** (% Correct)	**16 Zones** (% Correct)	**100 Zones** (% Correct)
Full-Scale	100%	100%	100%
Quarter-Scale (Pre-Compression)	63.3%	83.3%	85.0%
Sixteenth-Scale (Pre-Compression)	58.3%	60.0%	63.3%
Quarter-Scale (Post-Compression)	86.7%	100%	100%
Sixteenth-Scale (Post-Compression)	80.0%	98.3%	100.0%

Table 9.2: Raspberry Pi Camera dataset classification performance.

Compression Factor	**Single Zone** (msec)	**16 Zones** (msec)	**100 Zones** (msec)
Full-Scale	505.6	511.7	562.7
Quarter-Scale	171.9	185.6	207.8
Sixteenth-Scale	40.4	50.1	62.2

Table 9.3: Compressed camera source-identification run-time.

The run-time and memory overheads for compressed camera source identification are given in Tables 9.3 and 9.4. This is first organized by zonal structure: single zone, 16 zones and 100 zones. The other variable is the compression factor: full-scale, quarter-scale and sixteenth-scale. The goal of this evaluation is to identify the optimal algorithm as a function of run-time. In generality compression significantly reduces run-time while number of zones incremental increases run-time. Note these metrics are purely for classification, and do not take into account the resources associated with compressing the image.

As expected, there are significant benefits to both metrics when using the compression methodology. These results indicate that the optimal condition for source identification is to use sixteenth-scale down-sampling with 16 zones. On the target hardware, this decreases run-time from 505.6 msec to 50.1 msec, a relative reduction of 91.1%, and memory from 115.33 Mb to 8.24 Mb, a relative reduction of 92.8%. These run-time reductions are accomplished with perfect test-accuracy. This makes it easy to justify this methodology for real-applications (factoring both CPU and memory needs).

The key takeaway is this methodology is able to retain relevant PRNU features when compressed. This is accomplished by using an intelligent sequence of extracting the PRNU enrollment and utilizing zonal analysis to add local sensitivity. This results in significant run-time and memory benefits at inference. This accomplishment is particularly relevant when addressing photo-realistic tampering. These attacks are designed to be imperceptible; as such, detection methods must be sufficiently sensitive (while meeting application latency requirements).

Compression Factor	**Single Zone** (Mb)	**16 Zones** (Mb)	**100 Zones** (Mb)
Full-Scale	115.33	115.33	115.33
Quarter-Scale	32.96	32.96	32.96
Sixteenth-Scale	8.24	8.24	8.24

Table 9.4: Compressed camera source-identification memory utilization.

9.2.6 Exp 3: Direct Face-Swap-Verification

This experiment evaluates where optimizations can be done for FR by applying FZV. I.e., the tampering-zone is only calculated on the zone containing the face's centroid. Tampering is done on the Ras Pi Camera data-set, utilizing the same cameras to acquire 50 photos containing face-swap volunteers (per camera). Swaps are conducted in manual and artificial intelligence fashions. Manual swap is conducted by identifying the boundary of the detected face in both images, resizing the proposed face to match the existing one and applying swap (no photo-realistic blending). The artificial intelligence approach employs a well-known online tool, Reflect (71). This approach intelligently uses facial-landmarks to align and merge faces together (also applying photo-realistic blending). This process is visualized using two test subjects in Fig. 9.1.

Face-swaps are done using faces from both the same camera and different cameras. The intention is designed to evaluate swap-detection sensitivity. While most scenarios would necessarily involve different cameras (e.g., acquiring an image from social-media and injecting it into the FR system), it is an interesting investigation. This approach generates 20 swaps per swap use case, sampling rate and zone. This is a total of 1,440 tampered images generated. For reference, the face-swap regions range from $\frac{1}{20}$ to $\frac{1}{16}$ image width (generally small). This data-set is designated Ras Pi - Face Swap.

The FZV algorithm is benchmarked on the Ras Pi - Face-swap data-set (1,440 images constructed from no-swap, manual-swap and AI-face-swaps (71)). The FZV tampering-score is evaluated only on the zone containing the face-centroid, using 1, 16 and 100 zones. Face-centroid-zone is estimated using the Matlab cascade object detector (122). Note that Matlab is used just for convenience; in deployment this would be the FR's detection algorithm. This should not effect impact algorithm performance.

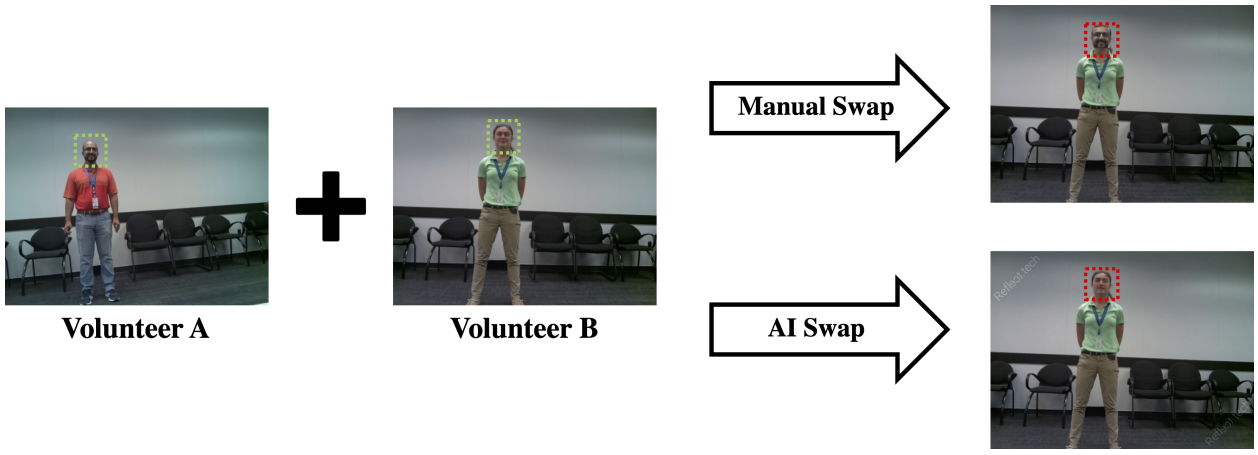


Figure 9.1: Visualizing face-swap-attacks methodologies.

9.2.6.1 Exp 3: Results

The FSA benchmark results are given in Table 9.5. Authentic (no swap) image is represented by "Authentic." For readability, the results table is simplified to represent all manual face-swap-attacks as "Manual Face-Swap" and all AI face-swap-attacks as "AI Face-Swap." This demonstrate the FZV approach generally outperforms the open-source algorithms for mitigating face-swap-attacks. In particular, the FZV algorithm is optimized at quarter-scale with 16 sub-zones (indicated by the †). Counter to intuition, the hypothesis of isolating the relevant pixels for noise analysis seems to reduce performance. The small number of pixels instead seems to produce a less reliable PRNU measurement, impacting tampering-score precision. One possible way to improve performance with reasonable overhead is to apply a nearest neighbor approach, including all adjacent zones.

The open-source results reinforce that the AI-face-swap detection is a significant challenge. At full-scale resolution, all three perform well. However, the aggressive calibration becomes problematic for down-sampled imagery. The deep learning approach (ELA CNN) (82) degrades to approximately a coin flip; the two deterministic algorithms (79; 78) effectively determined every image to be tampered (arguably worse). This shows noise-verification approach's utility.

Image Tampering	**FZV (1x1)**	**FZV (4x4)**	**FZV (10x10)**	**ELA CNN**	**Forgery Tool**	**DWT**
Authentic FS	100%	100%	100%	91.3%	98.8%	26.2%
Manual Face-Swap-Attack FS	100%	100%	100%	100%	99.0%	100%
AI Face-Swap-Attack FS	100%	100%	100%	100%	99.0%	100%
FS Mean	**100%**	**100%**	**100%**	**98.3%**	**99.8%**	**85.3%**
Authentic QS	92.5%	92.5%	92.5%	48.8%	30.4%	1.3%
Manual Face-Swap-Attack QS	100%	100%	92.5%	57.5%	100%	100%
AI Face-Swap-Attack QS	91.3%	87.5%	45.0%	53.8%	100%	98.8%
QS Mean	**95.0%**	**93.5%** †	**73.5%**	**54.3%**	**74.8%**	**79.8%**
Authentic SS	85.0%	85.0%	85.0%	45.0%	0.0%	0.0%
Manual Face-Swap-Attack SS	81.8%	81.8%	77.5%	51.8%	100%	100%
AI Face-Swap-Attack SS	81.3%	81.3%	27.5%	51.3%	100%	94.4%
SS Mean	**82.0%**	**82.0%**	**59.0%**	**50.2%**	**80.0%**	**77.8%**

Table 9.5: Face-swap-attack verification Results. This table gives the benchmarking results for the FZV, ELA CNN, Forgery Tool and DWT algorithms for FSA detection. The FZV zone distribution is denoted for each given column. For space purposes full-scale is abbreviated to FS; quarter-scale is abbreviated to QS and sixteenth-scale is abbreviated to SS.

Image Tampering	FZV (1x1)	FZV (4x4)	FZV (10x10)	ELA CNN	Forgery Tool	DWT
Face-Swap-Attack Full-Scale	384.1	40.0	14.7	247.0	207.9	48.2
Face-Swap-Attack Quarter-Scale	90.2	4.6 †	2.0	206.4	67.3	9.1
Face-Swap-Attack Sixteenth-Scale	22.8	2.0	1.0	181.3	4.7	5.7

Table 9.6: Face-swap-attack verification run-time. This table gives the run-time results for the FZV, ELA CNN, Forgery Tool and DWT algorithms for SDA detection. The FZV zone distribution is denoted for each given column. Units are in milliseconds.

In all cases, it is difficult to detect the compressed AI face-swaps. Two hypotheses are proposed. First, fewer pixels are tampered due to the intelligent blending. Second, the landmark-merged-face does not resemble the target enough to fool a reasonable face recognizer. The first hypothesis is challenging to mitigate; however, the second implies the identification algorithm can provide end-to-end security. Note that the actual identification would be done at full-scale resolution. The benefit of the noise-verification-framework is to only perform the authenticity analysis at compression; this means the full precision of the identification algorithm can be utilized.

The FSA run-time benchmark results are given in Table 9.6. The optimized algorithm is indicated by †. These results demonstrate the noise-verification-framework can provide a notable advantage in efficiency. When it comes to face-swap-attack mitigation, the optimal open-source algorithm would be DWT at full-scale. This takes 48.2 msec, in comparison to the ZEF taking 4.6 msec. This shows a substantial improvement, where the ZEF approach is ideal for doing per-frame analysis.

9.2.7 Exp 4: Face-Recognition Evaluation on Tampered Imagery

The FSA can be analyzed from an end-to-end perspective, factoring in the actual recognition rates. To evaluate this, a FR algorithm is constructed to verify identity. This algorithm is intentionally simplistic and is constructed using Matlab's face detector, histogram-of-oriented-gradients features and a support-vector-machine trained on the data-set participants.

9.2.7.1 Exp 4: Results

The face-swap recognition results are given in Table 9.7 (see next page). The results demonstrate that swaps capable of spoofing the PRNU tampering-score approach are not accepted by the facial recognition model. This inherently helps mitigate the risks of low-resolution AI-face-swaps not being detectable.

Swap Method	Face Classification Rate (%)		
	Full-Scale	Quarter-Scale	Sixteenth-Scale
Control	100%	100%	97.0%
Swap Same Camera - Manual	100%	100%	72.5%
Swap Different Camera - Manual	100%	100%	65.0%
Swap Same Camera - AI	0.0%	0.0%	20.0%
Swap Different Camera - AI	0.0%	0.0%	35.0%

Table 9.7: Face-swap identification accuracy. The AI swaps clearly degrade accuracy.

It is relevant to note this experiment is slightly flawed by employing a simplistic face-recognition algorithm. The theory is designed to indicate a simple algorithm can reject compressed swaps, but more sophisticated algorithms may actually be more capable at handling the swapping-artifacts. This is potentially observed with FaceNet being fooled by face-swap-attacks (26). Regardless, the findings validate that mitigating the FSA can be an end-to-end solution.

9.2.8 Exp 5: Simulated Service-Denial-Verification

After verifying that compressed PRNU can sufficiently verify face-integrity, Exp 5 evaluates a simulation of service-denial, where faces are removed (i.e., there is no face to detect). Worst case analysis is conducted by tampering images from the Ras Pi Camera data-set.

Face-removal across the image is simulated by randomly swapping blobs using matching scenes across images. That is to say a blob of matching background is placed, mimicking the effect of removing faces. The swap shape chosen is a circle to mimic the shape of a face. A swap example is shown in Fig. 9.2 (red outlines shown only identify the swap region and are not actually present). The ZEV algorithm is evaluated on the Ras Pi - General Swap data-set using 1, 16 and 100 zones.

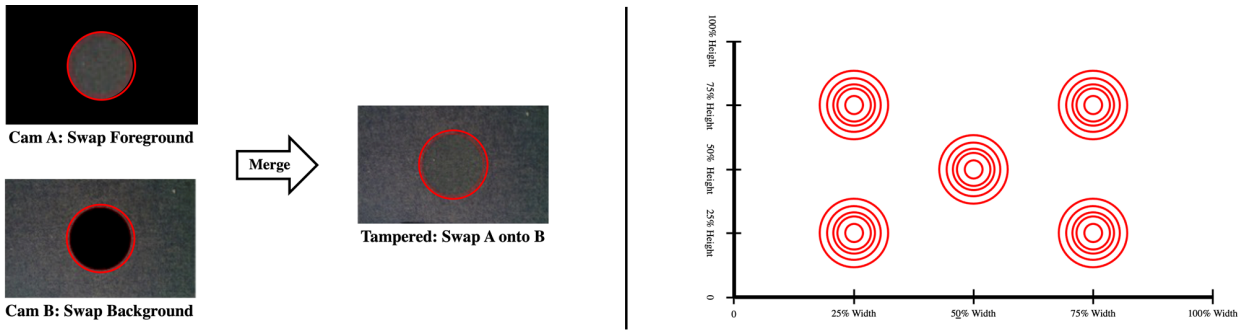


Figure 9.2: Face denial-of-service is simulated using blob swaps. Blobs are randomly exchanged across images to evaluate the tampering-score sensitivity to size and position.

Note that this is simply a tampering detection sensitivity analysis. It is not necessary to specifically remove objects in the scene, simply detect the tampering presence. One theoretical improvement would be to incorporate a deep-learning algorithm for photo-realism. This approach, however, is not employed as using such a tool would likely affect the whole image.

9.2.8.1 Exp 5: Results

The SDA benchmark results are given in Table 9.8. Authentic (no swap) images are represented as "Authentic." The results demonstrate the ZEV approach also generally outperforms the open-source algorithms for mitigating service-denial-attacks. The ZEV optimizes performance at sixteenth-scale with 100 sub-zones (indicated by the ‡). 100% accuracy is achieved over all blob-swaps and authentic images with relative efficiency. The utility of the zonal analysis really presents itself here, where significant compression can be applied while retaining full robustness.

The open-source algorithms conversely show significant degradation with slight compression. The deep-learning algorithm again seems to stabilize at a coin-flip, but generally speaking all are unreliable under heavy compression. There is at least less over-fitting this time, showing a smaller difference between authentic and blob-swap images. This is postulated to be a result of the very small swaps, where the lack of local analysis significantly deteriorates performances. This further validates the utility of the proposed-framework when using compression.

Image Tampering	**FZV (1x1)**	**FZV (4x4)**	**FZV (10x10)**	**ELA CNN**	**Forgery Tool**	**DWT**
Authentic FS	100%	100%	100%	99.7%	99.0%	70.2%
Service-Denial-Attack FS	100%	100%	100%	99.8%	99.0%	41.3%
FS Mean	**100%**	**100%**	**100%**	**99.8%**	**99.0%**	**46.1%**
Authentic QS	92.5%	100%	100%	54.8%	90.3%	59.5%
Service-Denial-Attack QS	82.3%	98.8%	100%	64.7%	70.0%	34.9%
QS Mean	**84.0%**	**99.0%**	**100%**	**63.0%**	**73.3%**	**39.0%**
Authentic SS	85.0%	92.5%	100%	40.2%	26.1%	55.0%
Service-Denial-Attack SS	74.4%	89.1%	100%	59.9%	0.1%	16.0%
SS Mean	**74.4%**	**89.0%**	**100% ‡**	**56.6%**	**12.7%**	**9.3%**

Table 9.8: Simulated service-denial-attack verification results. This table gives the benchmarking results for the FZV, ELA CNN, Forgery Tool and DWT algorithms for SDA detection. The FZV zone distribution is denoted for each given column. For space purposes full-scale is abbreviated to FS; quarter-scale is abbreviated to QS and sixteenth-scale is abbreviated to SS.

Image Tampering	FZV (1x1)	FZV (4x4)	FZV (10x10)	ELA CNN	Forgery Tool	DWT
SDA Full-Scale	507.7	537.6	1139.6	225.6	193.8	46.3
SDA Quarter-Scale	97.3	103.9	214.3	192.6	67.9	9.9
SDA Sixteenth-Scale	33.9	50.2	106.9 ‡	168.8	4.8	5.3

Table 9.9: Service-denial-attack verification run-time. This table gives the run-time results for the FZV, ELA CNN, Forgery Tool and DWT algorithms for SDA detection. The FZV zone distribution is denoted for each given column. Units are in milliseconds.

The SDA run-time benchmark results are given in Table 9.9. The optimized algorithm is indicated by ‡. Service-denial-attack mitigation is inherently more computationally expensive. Rather than just verifying the detected face, the full image must be verified for authenticity (e.g., face-removal). Here the ZEV algorithm again shows notable advantages in efficiency. The optimal open-source algorithm is the Forgery Tool at full-scale, which takes 193.8 msec in compared to the ZEV's 106.9 msec. While an improvement, this is still too slow for per-frame verification. Instead, it is suggested to do a periodic full-image authenticity challenge (e.g., the start of FR service and every few seconds later). This can mitigate the SDA while minimizing computational overhead.

9.3 Conclusion

This chapter addresses photo-realistic FR tampering through a noise-verification framework. A tampering score is assessed by measuring deviation from an expected camera 'noiseprint", photo-response-non-uniformity (PRNU), in a zonal fashion. Experimental results demonstrate reliable integrity-verification, robust to attack size and location. This approach is also compressed to achieve imperceivable run-time.

Given this method employs compression, it is important to verify the PRNU features remain robust. To do this, a pair of source-identification experiments are conducted on public datasets. This is designed to evaluate feature sensitivity as a function of number of zones and compression factor. Evaluation results show that features remain robust even when down-sampling to sixteenth-scale so long as at least 16 zones are used. The key finding is that compression order matters when performing the enrollment. The template image PRNU should be first extracted and then compressed for future challenge verification; down-sampling the image then extracting PRNU does in fact remove relevant features.

The integrity-verification methodology is next evaluated on face-swap-attacks and service-denial-attacks. This methodology similarly enrolls the camera PRNU then evaluates a zone-based correlation score with future images to verify authenticity. Experimental results show the proposed framework is both robust and significantly faster than the benchmarked algorithms. The 16 zone, quarter-scale down-sampled algorithm is not only robust, but can verify facial-authenticity in under 5 msec on CPU. This is a robust contribution as it meets the goals of securing face-recognition in an imperceivable fashion. Detecting SDAs is inherently more computationally expensive as it requires scanning the full image. This can be robustly achieved in approximately 100 msec, suggesting that a periodic full-image verification is ideal to optimize security and user-experience.

It is recommended that future research focus on removing the need for apriori knowledge of the camera. While robust, this method would not work on applications that cannot enroll the source-camera (e.g., social-media platforms). In theory, the "noiseprint" deviations caused from swaps or blending images could be identified via anomaly detection methods. This approach would likely require deep-learning but would introduce value beyond face-recognition.

CHAPTER 10

Related Application: Distilling Facial-Structure with Teacher-Tasks

This dissertation demonstrates that noise-inspired multi-task-learning (MTL) can improve facial-liveliness-verification (FLV) performance. This concept performs rather robustly, but is not necessarily an obvious concept to try. The inspiration for utilizing auxiliary-tasks to address intra-class variations actually comes this related application. While presented last, it was actually conducted prior to much of the FLV research and served as a technical foundation. Hence, it is included in the dissertation for completeness.

Face-identification is the process of identifying the person from features within the face-crop. While a generally solved problem in comparison to FLV, there are still false-rejection issues when it comes to head-pose. Pose-variations are facial rotations over yaw and pitch. These change the relative-position of key-points (e.g., nose, eyes) and introduce variance within identity classes. As such, face-recognition (FR) algorithms can struggle to discern the same person rotating from different people (123). This can be particularly problematic when considering strict standards on false-acceptance-rate (1), where the aggressive thresholds can cause pose-variations to result in false-rejections (123).

Current state-of-the-art methods rely on alignment techniques and/or sophisticated loss-functions to address pose-variability. Alignment methods can be simple, such as landmarks-based warping (37), or as sophisticated as projecting onto a 3D mannequin and rotating to be cooperating (124; 125). These are designed to constrain the problem though introduce the risk of bias from the pre-processing algorithm. Furthermore, these methods often only work well for small yaw and pitch values, degrading notably with large poses. Loss methods conversely are aimed at better understanding intra-class variance. Contrastive methods (37) (which may include transformations (13)) can improve performance without modifying the face, though often also degrade with large poses. These methods are clearly valuable but insufficient. To date no algorithm has achieved 100% on the competition dataset, labelled-faces-in-the-wild (126), indicating that pose-robust FR is still an on-going challenge.

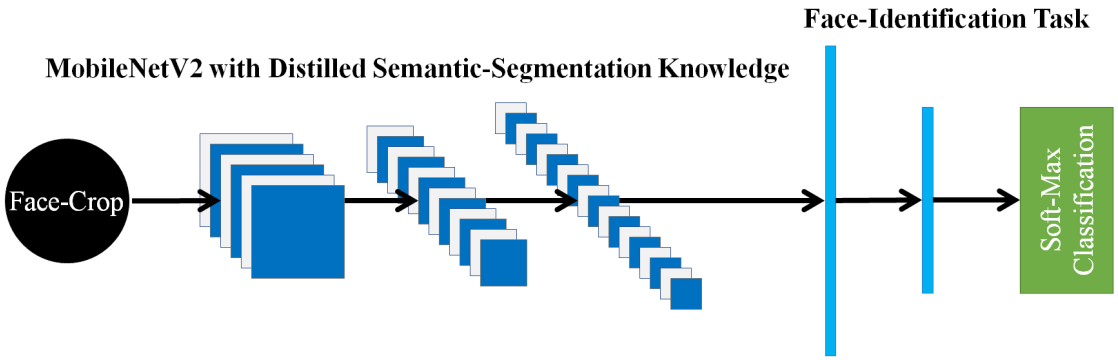


Figure 10.1: Seg-Distilled-ID network for pose-invariance.

This chapter proposes a human inspired approach to address pose-variations. Humans have a strong understanding of facial-structure, perceiving how the same person looks across poses. Intuitively, this semantic knowledge can be encoded into deep-learning (DL) algorithms, specifically using a novel MTL approach. This encoding is done by using semantic-segmentation as a teacher-task. I.e., it is used only for joint-training to learn facial-structure features, then "distilled" to improve run-time. This type of network is illustrated in Fig. 10.1. It is demonstrated that this approach not only generates best-in-class features, but can do so with a small parameter space. This validates the concept of using human-inspired tasks to improve algorithm features.

10.1 Distilling Knowledge with Teacher-Tasks

State-of-the-art face-identification algorithms often include alignment methods. Simplistic rectification using landmarks (e.g., the face is rotated based upon alignment of eyes and mouth) (12) is rather popular. This helps theoretically constrain the number of presentations possible, but comes with the potential risk of ironically distorting the face. If the face has notable yaw or pitch, it's rather possible the landmark-based warping can destructively effect identification accuracy. For this reason, best-in-class methods tend to do 3D projections. That is to say the face is first projected onto a mannequin, then rotated in 3D space to a cooperative perspective (125; 124).

These 3D projections, however, also degrade with notable yaw or pitch. This is hypothesized to a result of needing to infer much of the facial-features (which are obscured by the pose-rotation) from a mannequin. For this reason, some prefer to simply use a contrastive-loss function, such as triplet (37) or cosine (13), to implicitly observe the impact of pose-rotations. This approach is more consistent, but has not been shown to completely address the pose-variance sensitivity. As a point of discussion, it is acknowledged that they do intuitively make sense and can be considered to be used in combination with the proposed framework.

This research proposes that pose-sensitivity can be overcome by encoding the human-perspective. Humans observe the facial-structures and intuitively understand how this corresponds with identification. Hence, the proposal is to apply a novel application of multi-task-learning (MTL) by utilizing semantic-segmentation as a teacher-task. A network is jointly trained on facial-structure with identification, then "distills" the teacher-task. The "distilled" semantic-features enable the encoder to generate robust features, enabling efficient pose-invariance recognition. One of the major gaps in the reconstructive methods is that the projection and identification networks are independent, such that there is no knowledge-sharing. This approach directly addresses this by optimizing facial-structure features for identification.

10.1.1 Encoding Knowledge via Joint-Learning

Training knowledge is back-propagated into Neural-Networks through loss functions. Each time a training sample is inferred, the "distance" between the inferred value and training label is calculated and used to adjust the weights. Depending on the application a variety of loss functions can be used. For example, when estimating the position of key-points, Mean-Squared-Error is commonly used. This is the distance between inference y_i and label $\hat{y}_i$ for N_{MSE} values as given in (10.1).

$$MSE = \frac{1}{N_{MSE}} \sum_{i=1}^{N_{MSE}} (y_i - \hat{y}_i)^2 \tag{10.1}$$

Conversely, for classification problems a log-loss is used (i.e. Cross-Entropy). This better separates out the relative distance between classes by incorporating classification probability, p_{CE}, for M_{CE} classes as given in (10.2).

$$CE = -\frac{1}{M_{CE}} \sum_{i=1}^{M_{CE}} y_i \log p(y_i) \tag{10.2}$$

For biometric-joint-learning (BJL), these losses can be jointly combined in the form of an expected value. The expected value of each task loss, $L_t(Y_t, \hat{Y}_t)$ (where Y_t and $\hat{Y}_t$ are the task inference and label vectors), is given in (10.3). This expected value incorporates loss weights w_t and normalization η_t .

$$Loss = \sum_{t=1}^{K} w_t \eta_t \times L_t(Y_t, \hat{Y}_t) \tag{10.3}$$

For constructive training the weights need to both reflect the task complexity and the loss scale. Difficult tasks need extra weight to avoid training early stopping. Loss scale normalization helps ensure the information is back-propagated with the same intensity. For example, binary-cross-entropy will have loss on a 0 to 1 scale, whereas Mean-Squared-Error is proportional to the number of values square; these should both be normalized to the 0,1 range. Improper weighting and normalization can otherwise lead to stopping early or even destructive interference when training.

10.1.2 Seg-Distilled-ID Network

The proposed Seg-Distilled-ID network is shown in Fig. 10.2. The segmentation-task functions as a teacher, helping the ID-task better converge towards optimal weights. The teacher-task is removed once training is complete.

The network assumes a U-Net architecture (127). U-Net is selected both for its applications to biomedical semantic-segmentation (127) and option for efficient MobileNetV2 encoder (108). A MobileNetV2 backbone (108) encodes features for parallel identification and semantic-segmentation tasks. The identification-task is constructed by applying a global-average pooling layer, followed with a dense, 128-neuron, feature layer (ReLU activation (128)) and a dense, 67-neuron, classification layer (soft-max activation (129)). The segmentation-task is constructed using the Pix2Pix decoding layers (130) (e.g. final segmentation output of 128 by 128).

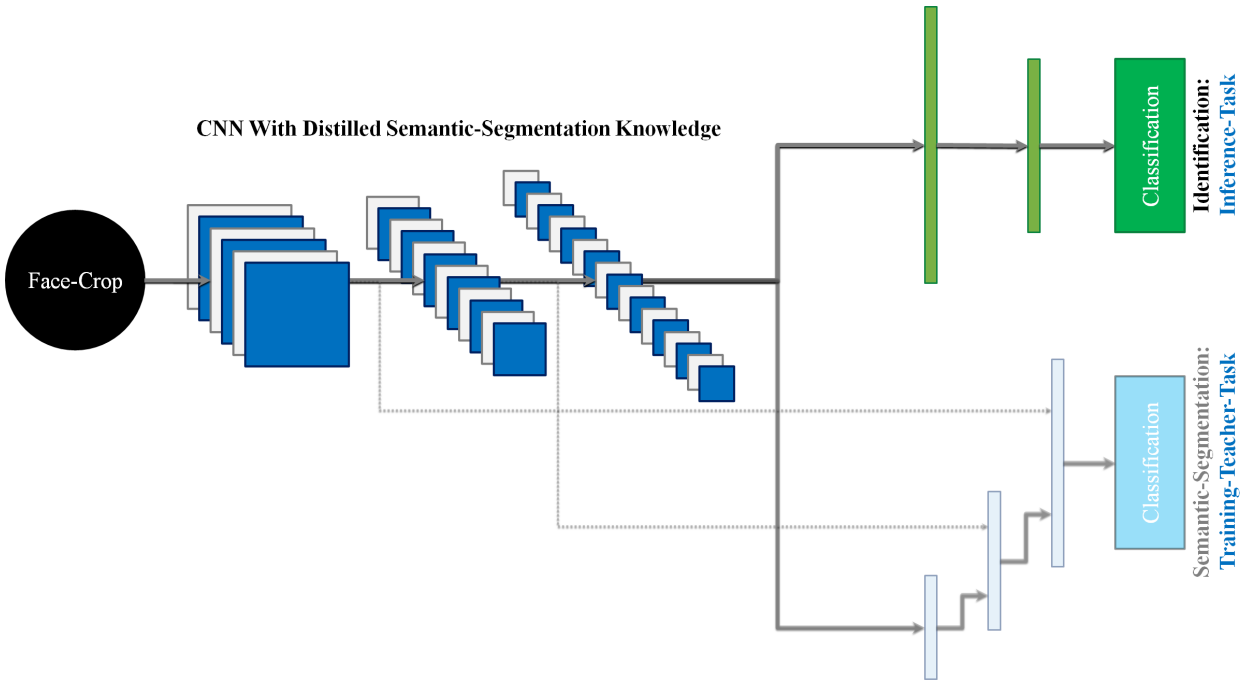


Figure 10.2: Improving face-identification with semantic-segmentation teacher.

Both tasks use a categorical-cross-entropy loss, as shown in (10.2). This better separates out the (log) distance between classes by incorporating probability of the observation, o, belonging to the label-class, c. This probability can be defined as $p(o, c)$ (129). A binary label, $\hat{y}$, indicates whether the prediction matches the correct class. This is done per class c of M (129) in an expected-value fashion.

Removing the teacher-task after training significantly reduces the network inference parameters. This drop from 6.5M to 2.4M is approximately a 63% reduction, most of which are fully-connected layers. I.e., the parameter reduction should strongly correlate with run-time reduction (fully-connected-layers are harder to parallelize than convolutional-layers). The purpose of this architecture is to demonstrate that label precision is ultimately key for generating robust features, and can even be utilized with an efficient encoder. The final inference network structured is illustrated in Fig. 10.1, where the encoder color change represents the segmentation knowledge-distillation (see start of chapter).

10.2 Performance Evaluation

The objective is to determine which network structures can generate the best features for pose-invariant FR. The proposed method is an MTL approach to generating features. The idea is to concurrently learn tasks that describe faces regardless of orientation, then remove the teacher to improve run-time. This needs to be validated against the traditional state-of-the-art feature encoders. This is done by comparing the MobileNetV2 (108) with and without the teacher-task, as evaluating three superior feature encoders.

Note the loss function is also a relevant feature-encoding tool. In principle contrastive methods, such as triple (37) and cosine (13), do improve identification accuracy. In practice these are difficult to implement with MTL. Hence, only categorical loss is used for consistency. An evaluation comparing MTL encoding versus loss methods (or even potentially combining the two) would be a relevant next step.

10.2.1 Validation Algorithms

The teacher-task network is benchmarked against three state-of-the-art encoders and MobileNetV2 without teacher-task (108). A comparison of other MTL methods is desired but not possible without the correct annotations. For example, Yin et al demonstrate learning head-pose with identification can improve performance (131), but the Mut1ny data-set does not contain the same pose-annotations.

Each benchmark network follows the same ID task-structure. That is to say an encoder generates the features, where are global-average-pooled, then classified using a 128-neuron dense feature-layer (ReLU activation) (128) and 67-neuron dense ID-classification-layer (soft-max activation) (129). The following network feature-encoders are used:

1. MobileNetV2 (108)
2. ResNet-101 (132)
3. VGG-19 (133)
4. IncepvtionV3 (109)

Each network is referred to as the encoder "-ID". E.g., validation-network 1 is designated "MobileNetV2-ID." All feature-encoders come pre-trained on ImageNet (110). Networks are compiled and trained in the same fashion, up to 125 epochs with a validation-loss patience of 20. Due to space constraints, training and validation curves are not shown.

10.2.2 Research Limitations

MTL is a new avenue of the DL field. At the time of this research, minimal art is found on other joint-learning applications for facial-identification. From this perspective this research does not have a true validation benchmark except to compare the network performance with and without the contextual tasks. Secondary validation is then done by contrasting performance against general state-of-the-art methods.

10.2.3 Exp 1: Pose-Invariant Identification

This experiment evaluates identification performance under high pose-variation. The Mut1ny Face/Head Segmentation (commercial edition) data-set (134) is used, employing 67 synthetic users with 150-250 unique perspectives (pose and background) each (11830 total). Each face is annotated with 14 structure classes: lips, left-eye, right-eye, nose, skin, hair, left-eyebrow, right-eyebrow, left-ear, right-ear, teeth, facial-hair, spectacles and background. These are cropped using the Dlib face detection tool (135). Model verification-accuracy is measured following Labelled Faces in the Wild procedures (126). Each person has 90% of their face-perspectives associated for training (8,320) and validation (2,080); test accuracy is evaluated on the remaining 10% (1,430).

Sample images from the Mut1ny dataset (along with segmentation-masks) are shown in Fig. 10.3 (see next page). While there are only 67 people, it is a very challenging face-recognition data-set. There are differences in pose, accessories, facial-hair and illumination. These significantly increase *intra*-class variability.

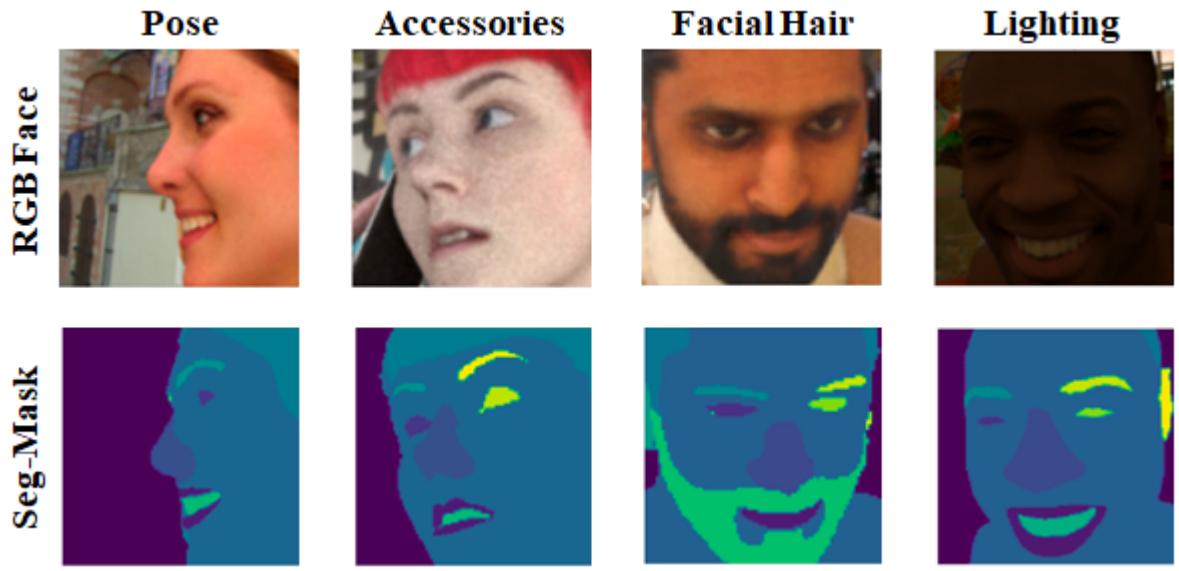


Figure 10.3: Mut1ny dataset challenging image samples. Observe how the semantic-segmentation masks better visualize the relevant features in the faces.

10.2.3.1 Exp 1 Results

The pose-invariant benchmarking results are shown in Table-10.1. As generally expected, having a stronger encoder correlates with better ID classification. All networks but MobileNetV2-ID train to a validation accuracy of at least 95% (training data not shown for space). This is intuitively understandable. For example, the MobileNetV2 architecture is designed to be as compact as possible (as is surprisingly effective on ImageNet (108)). Conversely, the ResNet101 (132) and InceptionV3 (109) architectures are designed to maximize feature projection through depth wise and breadth wise convolutions respectively. One would naturally assume the better encoders are more accurate for identification; the real question is whether the teacher-task framework can overcome this?

Network	Parameters	Test Accuracy
MobileNetV2-ID	2.4M	21.9%
ResNet-101-ID	43M	81.6%
VGG-19-ID	20M	96.1%
InceptionV3-ID	22M	96.3%
Seg-Distilled-ID	**2.4M (6.5M +Seg)**	**99.9%**

Table 10.1: Benchmarking the semantic-segmentation teacher-task framework on Mut1ny dataset. Note how the proposed Seg-Distilled-ID is not only the most accurate algorithm, but also is tied for fewest inference parameters.

The answer is clearly yes. Despite MobileNetV2-ID over-fitting, the Seg-Distilled-ID has the highest accuracy score evaluated. This is achieved while retaining the MobileNet architecture's efficiency (approximately one-tenth of the VGG and Inception network parameters). The parenthesis indicates that 2.4M parameters are used for inference and 6.5M are used for jointly training with the teacher-task. Recall again that the distilled training-parameters are fully-connected-layers, and therefore a sizable run-time reduction.

The parameter efficiency is explainable by using the semantic-segmentation knowledge to select optimal features. Top-tier encoders theoretically use the large parameter-spaces to implicitly infer context, enabling them to perceive information the base MobileNetV2-ID cannot. This methodology instead explicitly provides context through the semantic-segmentation teacher-task. Intuitively, these features better associate relevant facial components across poses for precises identification. This robustness enables the proposed Seg-Distilled-ID to efficiently achieve best-in-class performance.

Note that generalized pose robustness is very much novel. Others demonstrate re-aligning the face in 3D space can improve identification robustness (such as LDF-Net (125) and GridFace (124)). These methods are effective but degrade as yaw and pitch increase. This degradation is hypothesized to be a result of the 3D alignment algorithms synthetically inferring obscured facial components. This can potentially cascade bias from the alignment algorithm to the identification algorithm. The Seg-Distilled-ID avoids this bias by learning facial-structures in a one-shot approach.

10.2.4 Exp 2: Facial-Structure Feature Sensitivity

The results demonstrated in Exp 1 are outstanding, validating the teacher-task methodology. This is done by encoding facial-structure features using the semantic-segmentation task (14 semantic classes). This raises a fundamental question: which facial-structure features are most associated with identification accuracy? This is question is both academic and pragmatic in nature. There is academic value to being able to explain the results (which can inspire other related works), but more critically semantic-segmentation labelling is time consuming and expensive. The utility of this approach increases greatly if the annotation process can be simplified without degrading results.

Evaluating facial-structure sensitivity is done by merging annotation classes. If identification is sensitive to a given semantic feature, removing it from the teacher-task labels (e.g., merging it into background) should necessarily degrade results. Given there are 14 classes, a few intuitive approaches are taken to reduce the number of evaluation combinations. Features are associated by 3D impact (e.g., nose sticks out), temporal impact (e.g., eyes and mouth move) and symmetry

(e.g., relevance of left eye and right eye versus one eye class).

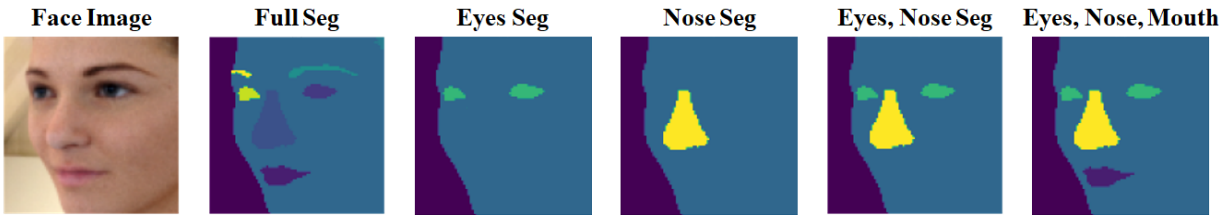


Figure 10.4: Facial-structure label merging visualization. Semantic-segmentation classes are merged to evaluate which facial-structures are most associated with identification.

Relevant samples from this process are visualized in Fig. 10.4. One can observe how the largest structure is clearly the foreground (which is not indicated as a unique class in the figure). The nose is the second largest structure. Intuitively, the nose should be critical both due to mass and the fact is varies the most with yaw and pitch. Following the intuition further, secondary structures that are smaller but vary temporarily (e.g., eyes and mouth) should also be useful as identification features. The goal is to identify the optimal feature set (considering accuracy and cost).

10.2.4.1 Exp 2 Results

The sensitivity analysis shows remarkable benefits can be achieved with just a few semantic classes. For starters, simply segmenting the face into foreground versus background shows a significant jump in identification accuracy. The next large jump is from including the nose. This intuitively makes sense, as the nose is the largest depth structure. Given it is largely centered on the face, it can act as a focus point for facial-orientation. These results then begin to outperform the top encoder from Exp 1 (InceptionV3) by including the eyes and mouth. These are temporal components, and necessarily will change regardless of the yaw and pitch. These results are given in Table 10.2 (see next page). All algorithms that outperform InceptionV3 are indicated by the *. Optimized teacher-task combinations are indicated by † and ‡, associating with efficient and robust annotation respectively.

The optimal efficient teacher-task combination is deemed to be combination of foreground, eyes, nose, mouth and background (indicated by the †). This combination yields effectively the same performance as the full 14 semantic classes with a fraction of the annotation cost (5 versus 14). An argument can be made that the optimal combination should include hair and glasses labels, as that is only one more class and does maximize the performance. This argument is furthered by the fact that these are temporal factors in the real-world, where people do change their appearance in these ways. The downside is annotating hair is extremely time consuming and therefore costly. For this reason, this group is designated the optimal robust combination (indicated by the ‡) as the

Encoder	Teacher-Tasks	ID Accuracy
MobileNetV2	None	21.9%
MobileNetV2	Face, BG	72.0%
MobileNetV2	Face, Eyes, BG	75.0%
MobileNetV2	Face, Nose, BG	93.2%
MobileNetV2	Face, Eyes, Nose, BG	96.0%
MobileNetV2	Face, Eyes, Hair, Glasses, BG *	97.9%
MobileNetV2	Face, Eyes, Nose, Mouth, BG *†	99.6%
MobileNetV2	Face, Eyes, Mouth, Hair, Glasses, BG *	99.8%
MobileNetV2	Face, Eyes, Nose, Mouth, Hair, BG *	99.9%
MobileNetV2	Face, Eyes, Nose, Mouth, Ears, Hair, Glasses, BG *‡	99.9%
MobileNetV2	Full 14 Semantic Classes *	99.9%

Table 10.2: Evaluating identification sensitivity to facial-structure features. Semantic-segmentation labels are merged into a general face foreground ("Face") or background ("BG") identify which classes are necessary for robustness. Algorithms that outperform the top encoder (InceptionV3) are indicated by the *. The optimal efficient subset is indicated by the † and optimal robust subset is indicated by the ‡.

benefits should present more on harder datasets, but is necessarily more challenging to annotate.

10.3 Conclusions

In conclusion, this chapter demonstrates the utility of incorporating teacher-tasks to optimize network features. Like traditional knowledge-distillation, the teacher-task concept is employs constructive learning with related tasks to encode contextual features, optimized for the primary task.

The first experiment utilizes semantic-segmentation to encode facial-structure features into identification networks. Benchmarking with state-of-the-art encoders shows the proposed Seg-Distilled-ID network achieves best-in-class performance using significantly fewer parameters. The knowledge transference is very strong, as the same network without teacher-task shows significant over-fitting. This verifies the hypothesis that including the teacher-task yields exceptional features - despite using a small parameter space.

The second experiment evaluates which facial-structure classes are actually relevant for identification. Interestingly, robust results can be achieved with just a few structures. The nose appears to be most critical and is theorized to function as a point of orientation. Temporal classes, such as eyes and mouth, then provide sufficient features to outperform the state-of-the-art validation algorithms. If the goal is to optimize annotation for efficiency, the recommendation is to include foreground, nose, eyes, mouth and background classes. If the goal is to optimize annotation for

robustness, the recommendation is additional include hair and glasses classes (as they also can be temporal features in the real-world).

The recommendation is to further evaluate this methodology with more complicated applications. The Mut1ny data-set (134) has only 67 subjects in the synthetic-face repository at this time. A pragmatic next step would be to apply this technique onto a competition dataset, such as Labelled Faces in the Wild (LFW) (126). LFW has significantly more people and is a good opportunity to evaluate more sophisticated network topologies (e.g., DeepLabV3 encoder (136) and contrastive identification loss (37)). Note, however, LFW does not have semantic-segmentation labels. This means either the Mut1ny features would need to be transfer learned or the dataset will need to be annotated.

Additionally, it would be rather interesting to evaluate the utility of semantic-segmentation features for facial-liveliness-verification. As introduced in the spectroscopy methodology, the material albedo can impact the reflectance distribution (see: chapter-6.1). This may be particularly relevant when considering more complicated spoofs, such as highly-realistic 3D masks, where the geometry alone may be insufficient.

CHAPTER 11

Conclusions

In conclusion, this dissertation successfully addresses monocular, single-frame facial-liveliness-verification (FLV). Revisiting the threat-model, there are two fundamental vulnerabilities: physical-spoof-attacks (PSAs) and face-swap-attacks (FSAs). These attacks are potent because of their ease of construction and efficacy. This research successfully mitigates both in a fashion is reliable and computationally efficient. This has translated into a series of academic publications and patents applications, which are presented in Appendices B and C respectively.

Addressing the PSA is a difficult task. Historical methods traditionally involve depth features from 3D sensors or temporal networks. The proposed material-spectroscopy approach takes inspiration from this, but achieves the desired goals by analyzing near-infrared reflectance-patterns. The National Institute of Standards and Technology recommends considering spoofs inspired from 2D images (14). These attacks (pictures, videos and simple-masks) lack the facial-complexity of a live-face. This yields less texture within the reflectance-patterns, a hypothesis that is modelled mathematically and verified empirically on a large-scale collected dataset. This collection is a key contribution, and includes 80,000 unique frames generated from 30 diverse adults under varied liveliness presentations, head poses, positions and lighting conditions. The mathematical model and validation are presented in Chapter-6.

One potential risk of the spectroscopy approach is image-noises that effect texture. The mathematical model assumes high-quality imagery with diffuse ambient. Real-world scenarios, however, will often introduce camera and environmental noises. This research pragmatically addresses these noises through synthetic noise-augmentation generators are designed. These are semi-realistic generators; they perturb the feature-space in a physics-informed fashion, but are not photo-realistic. While real data is ideal, repeating the entire collection under noise would be roughly a 13-fold increase in imaging. Initially, the spectroscopy algorithms are sensitive to the noises. This is quickly mitigated, however, by including noise-augmentations in the training process. This training approach improves noisy performance with no risk to clean data performance. Furthermore, noise-augmentations can potentially be used as a replacement for fully-contrastive data (reducing collection efforts). The noise-generators methodology and validation are presented in Chapter-7.

The efficacy of training with noise-augmentations implies there is value to learning noise-context. This intuition is quite literally applied next using a novel, noise-based multi-task-learning (MTL) approach. The auxiliary-noise-task (ANT) framework proposes that jointly learning liveliness and noise tasks will encode orthogonal features (for signal and noise respectively). This acts as a de-noising filter, where the classification tasks can better isolate the relevant features. Learning orthogonal tasks, however, requires a new training-loss paradigm. The task loss weights need to be dynamically re-balanced (as a function of validation loss) to ensure all tasks converge. Additionally, the tasks should be individually fine-tuned once the encoder is frozen. This process results in exceptional performance. The ANT networks trained on the fully-contrastive dataset essentially have perfect accuracy. This feature robustness is further demonstrated when using partially-contrastive data, where the ANT networks actually are more accurate than the baseline networks trained on fully-contrastive data. This is an exceptional contribution, as it both improves performance while reducing data-collection needs. The ANT methodology, experimentation design and results conclude the PSA assessment in Chapter-8.

The last FLV framework addresses photo-realistic image tampering. The primary concern is the FSA, where attackers digitally swap in the face of an enrolled user for authentication. This attack is addressed through a camera-noise verification strategy. Rather than identify the type of image transformation, image-integrity can be quickly verified against an enrolled noise profile. This can be optimized for run-time by applying compression and only verifying the facial-pixels. The experimentation shows the proposal is both more robust and faster than benchmarked open-source validation algorithms. The recommendation is to verify the face-pixels for noise-tampering on a per-frame basis. If service-denial-attack is also a concern (where attackers digitally remove faces), a periodic full-image-verification can also be done. The image-integrity verification methodology and validation are presented in Chapter-E.

For completeness, a related-application is presented for encoding facial-structure features. Humans intuitively learn to identify people from their facial-structures; hence, it is hypothesized teaching algorithms these structures can improve identification. The use-case considered is identification under notable pose and environmental variations (selected based off the availability of semantic-segmentation annotations). Another novel MTL framework is proposed using teacher-tasks; i.e., a second task is appended only for constructively improving the feature-space, then removed prior to inference. Here semantic-segmentation "teaches" the identification network. Experimentation shows the methodology generates notably better identification features than simply using more powerful networks. While this does not directly pertain to FLV, it is actually the inspiration for the ANT framework (it was completed first). The teacher-task methodology and validation are presented in Chapter-10.

11.1 Proposed Future Works

The fundamental FLV research is complete. If desired, there is opportunity to investigate the theory of related applications. For example, alternative-infrared-spectra (such as short-wave and long-wave infrared) have interesting perception capabilities. These are currently expensive, but have the potential to improve a multitude of biometric applications. Alternatively, both of the proposed MTL methodologies (auxiliary-noise-tasks and teacher-tasks) can be further explored. This can be done to either improve FR run-time by combining detection, identification or liveliness, or also address new biometric applications.

11.1.1 Alternative-Infrared-Spectra for Biometrics

Short-wave-infrared (SWIR) and long-wave-infrared (LWIR) offer the opportunity perceive deeper into the infrared spectrum. SWIR is a form of reflective light that is skin-tone invariant, but has high contrast with relevant spoof materials (paper, plastic, etc.). LWIR alternatively perceives thermal radiations. This also should theoretically provide very strong contrast across liveliness presentations, as live faces should have distinct patterns as a function of the vascular structure. In both cases this can improve feature reliability and can potentially employ simpler algorithms (faster run-time). A juxtaposition of liveliness-perspectives with these infrared technologies is visualized in in Fig. 11.1.

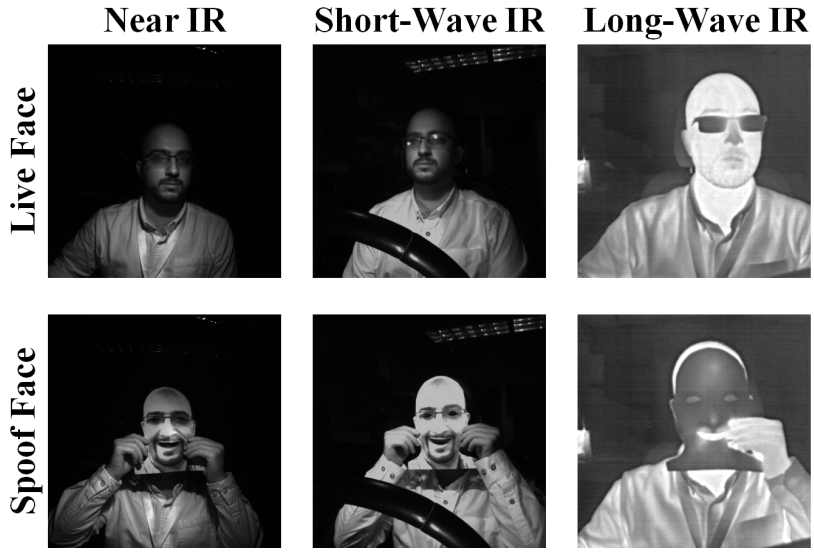


Figure 11.1: Liveliness presentations across various infrared spectra. Both short-wave and long-wave infrared improve liveliness contrast. Note how spoofs appear even brighter in SWIR and lack the necessary thermal radiation patterns in LWIR.

The proposed investigation is to identify which imaging technology offers the best utility across multiple biometrics applications. While the alternative infrared spectra do visually provide more liveliness contrast, it is sufficiently established that near-infrared can meet FLV requirements. Hence, the utility of these sensors is best evaluated in aggregate. The recommendation is to focus on applications that are particularly challenging with traditional NIR imaging. Relevant examples include:

- emotion-recognition
- drowsiness-recognition
- stress-recognition
- respiration-rate
- blood-pressure

Some of the biometrics are potentially solvable using temporal features. For example, emotion-recognition is difficult with single-frame algorithms, but can intuitively benefit from using time sequences. In this case the goal may be to simplify complexity (both improving reliability and run-time). Other biometrics likely require hyper-spectral imaging for basic robustness. For example, respiration-rate and blood-pressure are visually obvious in the thermal spectrum but still require temporal deep-learning networks.

Note that this infrared waveband comparison was originally included in the proposal. The work is academically interesting and there is a lot of untapped potential for SWIR and LWIR in the biometrics space. This said, investigating secondary hyper-spectral cameras is less valuable to the FLV than developing robust monocular, single-frame algorithms. For this reason, the proposal is modified to instead include the synthetic noise-augmentation and ANT frameworks.

11.1.2 Extending Multi-Task-Learning Frameworks

This dissertation presents two types of MTL frameworks. The ANT framework is about identifying noise presences to better isolate the primary task features. This can be done using either real or synthetic noise-augmentations (though for pragmatic purposes, synthetic noise is much easier to annotate). This framework is demonstrated to be extremely robust for FLV and should be investigated for other biometric applications as well. Apply ANT towards driver-state-monitoring would be particularly interesting when considering the societal value. As vehicle autonomy increases, there is a growing need to robustly characterize driver state occurs various noisy conditions. Driver-monitoring is already known to be affected by lighting effects; this methodology may naturally translate well.

The teacher-task framework is about utilizing a more complex task as a training teacher. The application presented here is for pose-invariant face-recognition. This is largely addressed for the Mut1ny dataset; as such, a natural next step is to evaluate the method on a competition dataset. The proposal is to first train a precise semantic-segmentation network as a super-annotator, and appropriately annotate a competition dataset such as Labelled-Faces-in-the-Wild. From there, the teacher-task framework would be applied to generate a new Seg-Distilled-ID network. This may include an evaluation of various encoders and including more sophisticated training-loss paradigms to maximize performance.

Additionally, it is relevant to evaluate the utility of semantic-segmentation features for FLV. This dissertation intentionally does not include highly-realistic 3D spoofs, but there is some niche value to addressing them. One thought process is to further explore material-reflectivity when the geometry is insufficient. This intuition can be accomplished by similarly encoding facial-structure features into the liveliness network.

Furthermore, it would be interesting to investigate combining liveliness and detection into a single network. This has the opportunity to both reduce run-time by combining encoders, and provide the greater context from the full-scene (e.g., observing someone holding up a picture or display). This proposal can theoretically be done using the DeepLabV3 (136), which is already designed to do full-scene semantic-segmentation. An example of this hybrid network is illustrated in Fig. 11.2.

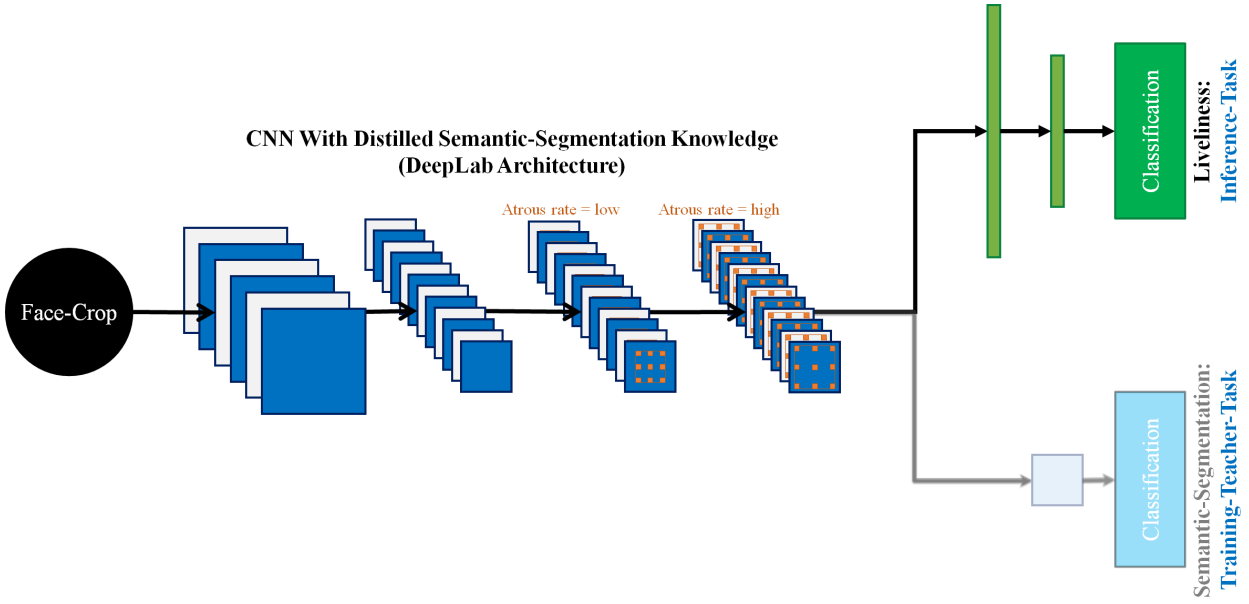


Figure 11.2: Incorporating liveliness and semantic-segmentation tasks at face-detector. The semantic-segmentation teacher-task with full-scene context may yield in strong performance and streamline pipeline. DeepLabV3 is potentially well suited for the application.

11.2 Final remarks

This concludes the facial-liveliness-verification research. The fundamental goals are achieved, demonstrating best-in-class performance for three attack vectors: physical-spoof-attack, face-swap-attack and service-denial-attack. These accomplishments are made possible by the substantial support from advisers, lab collaborators and sponsor Ford Motor Company. Their support is very much appreciated.

APPENDIX A

Glossary

This dissertation uses concepts from Computer-Vision, Machine-Learning and Cyber-Security. To improve readability, relevant terms are defined next.

A.1 Imaging-Technology

- **Camera** - imaging sensor built using photo-receivers.
- **CMOS** - Complementary metal-oxide-semiconductor, a type of photoreceiver material that is sensitive to light for Ultra-Violet, Visible and Near-Infrared.
- **Long-Wave-Infrared** - Waveband of light ranging from 8000nm to 1200nm, also known as thermal-infrared. Industry often uses the full waveband.
- **Near-Infrared** - Waveband of light ranging from 780 nm to 1000nm. Most common wavelengths used in industry are 850 nm and 940 nm.
- **Short-Wave-Infrared** - Waveband of light ranging from 1000nm to 1700 nm. Most common wavelength used in industry is 1400 nm.
- **Ultra-Violet** - Waveband of light ranging from 100 to 400nm. Most common wavelengths used in industry are 280 to 320nm range (UV-B) and 320 to 400nm (UV-A).
- **Visible** - Waveband of light ranging from 400 to 780nm. Most common wavelengths used in industry are 450 to 490nm range (blue), 520 to 560nm range (green) and 635 to 700 range (red).

A.2 Image-Processing

- **Local-Binary-Pattern** - Binary kernel for resolving local gradient information. Commonly used for texture features.
- **Histogram-of-Oriented-Gradients** - Image gradient orientation feature by slope magnitude and angle. Commonly used for object detection and classification features.
- **Optical-Flow** - Velocity of objects in the scene. Commonly used for 3D analysis, both for motion-planning and object classification.

A.3 Machine-Learning

- **Artificial-Neural-Network** - Multi-layer perceptron to project features into higher-dimensional space. Primarily used for harder problems where data is in series.
- **Class** - Each unique enumeration to be characterized.
- **Classification** - Process of determining the presented data's class from the provided features.
- **Convolutional-Neural-Network** - Advanced neural-network that uses convolutions to implicitly generate features. Primarily used for very hard image processing problems.
- **Deep-Learning** - Implicit feature generation using multi-layered neural-networks.
- **Feature** - Descriptor used for classification.
- **Margin** - Distance between class features. A measurement of separability.
- **Random-Forest** - Tree structure of randomly generated classifiers, where majority vote determines final classification. Primarily used for harder problems involving categorical data.
- **Support-Vector-Machine** - Binary classifier (a.k.a. perceptron) that estimates feature distance via hyper-planes (Mahalanobis distance). Primarily used for high-margin problems.

A.4 Cyber-Security

- **Enrollment** - Process of adding a user's face to the authentication database.
- **Authentication** - Process of validating a presented face against the enrollment database, and determining the identity (if enrolled).

- **False-Acceptance-Rate** - Rate of incorrectly classifying a non-enrolled challenge face as valid. This does not reflect imposters.
- **Spoof-Acceptance-Rate** - Rate of incorrectly classifying a spoof face challenge as valid.
- **False-Rejection-Rate** - Rate of incorrectly classifying an enrolled user's challenge face as not enrolled.

APPENDIX B

Academic Publications

The following publications are directly a result of this dissertation:

1. Ali Hassani, Jon Diedrich and Hafiz Malik. "Image Tampering Detection for Vehicle Systems." Ford Research Laboratory Technical Report. RLIS, 2019.

2. Ali Hassani and Hafiz Malik. "Securing Facial Recognition Systems: Spoof Vulnerabilities and Countermeasures." Biometric Technology Today. Elsevier, 2021.

3. Ali Hassani and Hafiz Malik. "Efficient Face-Swap-Verification Using PRNU." IEEE CDMA. 2022.

4. Ali Hassani, Hafiz Malik and Jon Diedrich. "Efficiently Mitigating Face-Swap-Attacks: Compressed-PRNU Verification with Sub-Zones." MDPI: Technologies. 2022.

5. Ali Hassani, Zaid El Shair, Rafi Ud Dual Refat and Hafiz Malik. "Semantic-Segmentation-Features For Pose-Invariant Face-Recognition." IEEE International Conference on Image Processing. 2022.

6. Ali Hassani, Jon Diedrich and Hafiz Malik. "Monocular Facial-Liveliness-Verification: Classifying Near-Infrared Reflectance Patterns." IEEE Transactions on Information Forensics and Security. Submitted.

7. Ali Hassani, Jon Diedrich and Hafiz Malik. "Addressing Monocular Facial-Liveliness-Verification Robustness To Camera and Environmental Noise." IEEE Transactions on Information Forensics and Security. To be submitted.

8. Ali Hassani and Hafiz Malik. "Facial-Liveliness-Verification With Auxiliary-Noise-Tasks: A Noise-Classification Framework To Optimize Features." IEEE Transactions on Information Forensics and Security. To be submitted.

APPENDIX C

Intellectual Property

This dissertation is funded by the Ford-UM Alliance Grant, Biometric Forensics. The following patents are jointly filed with Ford-Motor-Company (who owns exclusive licensing rights).

C.1 Image Integrity-Verification

The following patent applications are for verifying image authenticity.

1. *CAMERA IDENTIFICATION* (USPTO Case ID: **84215571US01**). A method for efficiently verifying camera source image authenticity using compressed, zonal "noiseprint" analysis. Camera source is calculated by zonal Peak Correlation Energy expected value against enrolled camera templates.

2. *CAMERA TAMPERING DETECTION* (USPTO Case ID: **84215575US01**). A method for efficiently verifying image authenticity using compressed, zonal "noiseprint" analysis. Tampering score is measured by deviation in zonal Peak Correlation Energy from the enrolled camera template.

3. *CAMERA TAMPERING DETECTION* (USPTO Case ID: **84215579US01**). A method for further optimizing image authenticity verification by only analyzing dected objects. Tampering score is calculated for only the zone containing the detected object's centroid.

4. *VISION SENSOR DYNAMIC WATERMARKING VIA NOISE CHARACTERIZATION* (USPTO Case ID: **84226056US01**). Dynamic watermarking encoded using camera noise as a means to be imperceptible.

C.2 Facial Anti-Spoofing

The following patent applications are for efficient anti-spoofing.

5. *COUNTERFEIT IMAGE DETECTION* (USPTO Case ID: **84238879US01**). Convenience facial authentication using Near-Infrared camera specular reflectance. Person is first identified, then verified their compensated specular reflectance meets the liveliness-enrollment-similarity score.

6. *COUNTERFEIT IMAGE DETECTION* (USPTO Case ID: **84227552US01**). Secure facial authentication with 2D and complex 3D mask anti-spoofing via co-registered CMOS and thermal cameras. CMOS camera is used to detect and identify the face; liveliness is determined using thermal analysis. System is secure with very efficient liveliness analysis.

7. *MATERIAL SPECTROSCOPY* (USPTO Case ID: **84279449US01**). Material source-identification using combined RGB-IR spectroscopy analysis. RGB provides material color context for Near-Infrared Material-Spectroscopy. This provides a naive Anti-Spoofing approach (versus specular-reflectance verification against enrollment).

8. *MATERIAL SPECTROSCOPY* (USPTO Case ID: **84279422US01**). Facial optical-tethering methods for Material-Spectroscopy liveliness-analysis. Facial distance and orientation are determined using deterministic key-points or Deep-Learning.

9. *MATERIAL SPECTROSCOPY* (USPTO Case ID: **84279413US01**). Facial environment-compensation methods for Material-Spectroscopy liveliness-analysis. Sequenced light toggling is used to detect the face with an illuminated frame and de-noise the background using non-illuminated frame analysis.

10. *MATERIAL SPECTROSCOPY* (USPTO Case ID: **84279409US01**). Facial segmentation methods for Material-Spectroscopy liveliness-analysis. In particular, emphasis is placed upon segmenting "skin" pixels either using deterministic key-points or semantically using Deep-Learning.

11. *SPOOF IMAGES FOR USER AUTHENTICATION* (USPTO Case ID: **84396269US01**). A method for utilizing generative adversarial networks to do a "spoof enrollment" to predict attack vectors.

12. *SEMANTIC SEGMENTATION* (USPTO Case ID: **84403055US02**). A multi-task biometrics network that utilizes an occupant monitoring camera to do identification, liveliness and other related biometric tasks.

13. *BIOMETRIC TASK NETWORK* (USPTO Case ID: **84403072US01**). A multi-task biometrics network that fuses semantic-segmentation task for Improved Liveliness and ID Tasks (Monocular Biometrics Network).

14. *BIOMETRIC TASK NETWORK* (USPTO Case ID: **84403086US01**). A multi-task biometrics network that fuses facial-landmarks task for improved emotion recognition and semantic-segmentation tasks.

APPENDIX D

Auxiliary-Noise-Task Network Sensitivity Analysis

The auxiliary-noise-task framework presented in Chapter-8 is designed to improve facial-liveliness-verification performance by explicitly learning noise-labels. The idea is that encoding the associated noise-features enables the deep-learning networks to better separate them from liveliness-features, and therefore improve performance.

Jointly learning signal and noise tasks, however, can be destructive. Hence, a sensitivity analysis is first performed to experimental identify the optimal network topology and training process. This appendix starts with presenting the sensitivity results for the material-spectroscopy algorithms.

The results given in Tables D.1, D.2 and D.3 show that the camera ANTs are most critical (see next pages). The environmental noises are generated using the camera noise generators to incorporate auto-exposure behaviors; intuitively it makes sense the camera noise-tasks could learn the necessary features. This said, performance is essentially identical when including the environmental noise tasks. While this dataset is not difficult enough, it is believed that using the environment ANTs would only generalization in the wild.

Encoder	MTL	Loss	Test	ACER	NPCER	APCER
Base Network (No Auxiliary-Noise-Tasks):						
MobileNetV2	None	Static	Clean	13.2%	20.8%	5.7%
MobileNetV2	None	Static	Noisy	22.2%	20.0%	24.4%
Camera Auxiliary-Noise-Task Networks:						
MobileNetV2	Soft	Static	Clean	1.3%	0.9%	1.8%
MobileNetV2	Soft	Static	Noisy	1.5%	2.3%	0.7%
MobileNetV2	Soft	Dyn	Clean	1.1%	1.0%	1.3%
MobileNetV2	Soft	Dyn	Noisy	3.2%	3.4%	2.9%
MobileNetV2 †	Soft	Dyn + Fine	Clean	0.8%	0.8%	0.8%
MobileNetV2 †	Soft	Dyn + Fine	Noisy	0.4%	0.7%	0.1%
MobileNetV2	Hard	Static	Clean	4.7%	4.3%	5.1%
MobileNetV2	Hard	Static	Noisy	6.0%	10.5%	1.5%
MobileNetV2	Hard	Dyn	Clean	0.9%	0.5%	1.2%
MobileNetV2	Hard	Dyn	Noisy	2.2%	1.7%	2.7%
MobileNetV2	Hard	Dyn + Fine	Clean	1.5%	0.8%	2.1%
MobileNetV2	Hard	Dyn + Fine	Noisy	1.8%	2.1%	1.5%

Table D.1: Evaluating optimal auxiliary-noise-task network topology: synthetic camera tasks. It is clear the ANT framework improves performance, where the best combination is indicated by the †. Dynamic scheduling is abbreviated to Dyn. Fine tuning is abbreviated to Fine.

Encoder	MTL	Loss	Test	ACER	NPCER	APCER
Base Network (No Auxiliary-Noise-Tasks):						
MobileNetV2	None	Static	Clean	13.2%	20.8%	5.7%
MobileNetV2	None	Static	Noisy	22.2%	20.0%	24.4%
Environmental Auxiliary-Noise-Task Networks:						
MobileNetV2	Soft	Static	Clean	0.5%	0.5%	0.5%
MobileNetV2	Soft	Static	Noisy	8.2%	2.6%	13.8%
MobileNetV2	Soft	Dyn	Clean	1.5%	1.2%	1.9%
MobileNetV2	Soft	Dyn	Noisy	2.4%	3.0%	1.8%
MobileNetV2 †	Soft	Dyn + Fine	Clean	2.8%	3.6%	2.1%
MobileNetV2 †	Soft	Dyn + Fine	Noisy	1.0%	1.3%	0.7%
MobileNetV2	Hard	Static	Clean	3.5%	1.3%	5.7%
MobileNetV2	Hard	Static	Noisy	3.0%	2.6%	3.4%
MobileNetV2	Hard	Dyn	Clean	0.8%	1.7%	0.0%
MobileNetV2	Hard	Dyn	Noisy	2.2%	2.4%	2.0%
MobileNetV2	Hard	Dyn + Fine	Clean	0.3%	0.0%	0.7%
MobileNetV2	Hard	Dyn + Fine	Noisy	1.4%	1.5%	1.3%

Table D.2: Evaluating optimal auxiliary-noise-task network topology: synthetic environmental tasks. It is clear the ANT framework improves performance, where the best combination is indicated by the †. Dynamic scheduling is abbreviated to Dyn. Fine tuning is abbreviated to Fine.

Encoder	MTL	Loss	Test	ACER	NPCER	APCER
Base Network (No Auxiliary-Noise-Tasks):						
MobileNetV2	None	Static	Clean	13.2%	20.8%	5.7%
MobileNetV2	None	Static	Noisy	22.2%	20.0%	24.4%
Camera and Environmental Auxiliary-Noise-Task Networks:						
MobileNetV2	Soft	Static	Clean	2.9%	5.1%	0.6%
MobileNetV2	Soft	Static	Noisy	6.0%	5.5%	6.5%
MobileNetV2	Soft	Dyn	Clean	1.1%	0.2%	2.1%
MobileNetV2	Soft	Dyn	Noisy	2.4%	3.1%	1.7%
MobileNetV2 †	Soft	Dyn + Fine	Clean	0.5%	0.8%	0.2%
MobileNetV2 †	Soft	Dyn + Fine	Noisy	0.9%	1.5%	0.2%
MobileNetV2	Hard	Static	Clean	3.4%	2.4%	4.5%
MobileNetV2	Hard	Static	Noisy	2.3%	3.5%	1.1%
MobileNetV2	Hard	Dyn	Clean	1.1%	1.9%	0.2%
MobileNetV2	Hard	Dyn	Noisy	1.9%	1.6%	2.2%
MobileNetV2	Hard	Dyn + Fine	Clean	4.2%	0.1%	8.2%
MobileNetV2	Hard	Dyn + Fine	Noisy	3.5%	4.1%	2.9%

Table D.3: Evaluating optimal auxiliary-noise-task network topology: synthetic camera and environmental tasks. It is clear the ANT framework improves performance, where the best combination is indicated by the †. Dynamic scheduling is abbreviated to Dyn. Fine tuning is abbreviated to Fine.

In addition to the fundamental material-spectroscopy algorithms, a new RGB-IR based occupant monitoring perspective as also considered. A second sensitivity analysis is conducted using a set of macro-descriptors, the results of which are given in Tables D.4 (this page) and D.5 (next page). This shows that when utilizing unstructured data, the covid-mask task is actually the most critical ANT though learning the identity is also helpful. Learning the identity intuitively masks sense as it enables the network to apply contexts of facial-structure and skin tone (which is known to change reflectivity). This said, there seems to be minimal benefit to using the robust (InceptionV3) network. This is promising, as the efficient network can be easily deployed on many real-time-systemss.

Encoders	**ANTs**	**MTL**	**Loss**	**ACER**	**NPCER**	**APCER**
MobileNetV2	None	None	Static	5.0%	5.0%	4.9%
MobileNetV2	ID	Soft	Static	3.5%	1.3%	5.7%
MobileNetV2	Covid	Soft	Static	3.0%	1.1%	4.9%
MobileNetV2	ID & Covid	Soft	Static	5.0%	2.7%	8.3%
MobileNetV2	ID	Soft	Dynamic	4.5%	2.0%	7.0%
MobileNetV2	Covid	Soft	Dynamic	2.9%	0.3%	5.6%
MobileNetV2 †	ID & Covid	Soft	Dynamic	2.6%	0.9%	4.3%
MobileNetV2	ID	Hard	Static	5.0%	2.6%	7.5%
MobileNetV2	Covid	Hard	Static	4.0%	2.3%	5.7%
MobileNetV2	ID & Covid	Hard	Static	4.0%	1.7%	6.3%
MobileNetV2	ID	Hard	Dynamic	5.1%	1.4%	8.8%
MobileNetV2	Covid	Hard	Dynamic	9.6%	1.1%	18.1%
MobileNetV2	ID & Covid	Hard	Dynamic	3.1%	%	4.7%

Table D.4: Auxiliary-noise-task framework liveliness results: efficient occupant-monitoring network (MobileNetV2). The best performing ANT topology is given by the †.

Encoders	**ANTs**	**MTL**	**Loss**	**ACER**	**NPCER**	**APCER**
InceptionV3	None	None	Static	5.0%	5.0%	4.9%
InceptionV3	ID	Soft	Static	3.5%	1.3%	5.7%
InceptionV3	Covid	Soft	Static	3.0%	1.1%	4.9%
InceptionV3	ID & Covid	Soft	Static	5.0%	2.7%	8.3%
InceptionV3	ID	Soft	Dynamic	4.5%	2.0%	7.0%
InceptionV3	Covid	Soft	Dynamic	2.9%	0.3%	5.6%
InceptionV3	ID & Covid	Soft	Dynamic	2.6%	0.9%	4.3%
InceptionV3	ID	Hard	Static	5.0%	2.6%	7.5%
InceptionV3	Covid	Hard	Static	4.0%	2.3%	5.7%
InceptionV3	ID & Covid	Hard	Static	4.0%	1.7%	6.3%
InceptionV3	ID	Hard	Dynamic	5.1%	1.4%	8.8%
InceptionV3	Covid	Hard	Dynamic	9.6%	1.1%	18.1%
InceptionV3	ID & Covid	Hard	Dynamic	3.1%	%	4.7%

Table D.5: Auxiliary-noise-task framework liveliness results: robust occupant-monitoring network (InceptionV3). The best performing ANT topology is given by the ‡.

APPENDIX E

Image-Integrity-Verification Sensitivity Analysis

The image-integrity-verification framework presented in Chapter-9 is designed to detect general tampering. For readability, only the optimized algorithms are shown there. This appendix expands those works with a full sensitivity analysis of integrity verification by size and location of tampering.

Image Tampering	**Single Zone** (% Correct)	**16 Zones** (% Correct)	**100 Zones** (% Correct)
Authentic Control	**100%**	**100%**	**100%**
Center $\frac{1}{12}$ Image	100%	100%	100%
Center $\frac{1}{16}$ Image	100%	100%	100%
Center $\frac{1}{20}$ Image	100%	100%	100%
Center $\frac{1}{25}$ Image	100%	100%	100%
Center $\frac{1}{50}$ Image	100%	100%	100%
Swap Center Mean	**100%**	**100%**	**100%**
Top Left $\frac{1}{12}$ Image	100%	100%	100%
Top Left $\frac{1}{16}$ Image	100%	100%	100%
Top Left $\frac{1}{20}$ Image	100%	100%	100%
Top Left $\frac{1}{25}$ Image	100%	100%	100%
Top Left $\frac{1}{50}$ Image	100%	100%	100%
Swap Top Left Mean	**100%**	**100%**	**100%**
Top Right $\frac{1}{12}$ Image	100%	100%	100%
Top Right $\frac{1}{16}$ Image	100%	100%	100%
Top Right $\frac{1}{20}$ Image	100%	100%	100%
Top Right $\frac{1}{25}$ Image	100%	100%	100%
Top Right $\frac{1}{50}$ Image	100%	100%	100%
Swap Top Right Mean	**100%**	**100%**	**100%**
Bottom Left $\frac{1}{12}$ Image	100%	100%	100%
Bottom Left $\frac{1}{16}$ Image	100%	100%	100%
Bottom Left $\frac{1}{20}$ Image	100%	100%	100%
Bottom Left $\frac{1}{25}$ Image	100%	100%	100%
Bottom Left $\frac{1}{50}$ Image	100%	100%	100%
Swap Bottom Left Mean	**100%**	**100%**	**100%**
Bottom Right $\frac{1}{12}$ Image	100%	100%	100%
Bottom Right $\frac{1}{16}$ Image	100%	100%	100%
Bottom Right $\frac{1}{20}$ Image	100%	100%	100%
Bottom Right $\frac{1}{25}$ Image	100%	100%	100%
Bottom Right $\frac{1}{50}$ Image	100%	100%	100%
Swap Bottom Right Mean	**100%**	**100%**	**100%**

Table E.1: General tampering detection sensitivity analysis - full-scale imagery.

Image Tampering	**Single Zone** (% Correct)	**16 Zones** (% Correct)	**100 Zones** (% Correct)
Authentic Control	**100%**	**100%**	**100%**
Center $\frac{1}{12}$ Image	90.0%	100%	100%
Center $\frac{1}{16}$ Image	80.0%	100%	100%
Center $\frac{1}{20}$ Image	80.0%	100%	100%
Center $\frac{1}{25}$ Image	80.0%	100%	100%
Center $\frac{1}{50}$ Image	80.0%	100%	100%
Swap Center Mean	**82.0%**	**100%**	**100%**
Top Left $\frac{1}{12}$ Image	90.0%	100%	100%
Top Left $\frac{1}{16}$ Image	80.0%	100%	100%
Top Left $\frac{1}{20}$ Image	90.0%	100%	100%
Top Left $\frac{1}{25}$ Image	80.0%	100%	100%
Top Left $\frac{1}{50}$ Image	80.0%	100%	100%
Swap Top Left Mean	**84.0%**	**100%**	**100%**
Top Right $\frac{1}{12}$ Image	85.0%	100%	100%
Top Right $\frac{1}{16}$ Image	80.0%	100%	100%
Top Right $\frac{1}{20}$ Image	85.0%	100%	100%
Top Right $\frac{1}{25}$ Image	75.0%	100%	100%
Top Right $\frac{1}{50}$ Image	80.0%	100%	100%
Swap Top Right Mean	**77.0%**	**100%**	**100%**
Bottom Left $\frac{1}{12}$ Image	75.0%	100%	100%
Bottom Left $\frac{1}{16}$ Image	85.0%	100%	100%
Bottom Left $\frac{1}{20}$ Image	80.0%	80.0%	100%
Bottom Left $\frac{1}{25}$ Image	80.0%	100%	100%
Bottom Left $\frac{1}{50}$ Image	65.0%	100%	100%
Swap Bottom Left Mean	**77.0%**	**96.0%**	**100%**
Bottom Right $\frac{1}{12}$ Image	95.0%	100%	100%
Bottom Right $\frac{1}{16}$ Image	75.0%	100%	100%
Bottom Right $\frac{1}{20}$ Image	75.0%	100%	100%
Bottom Right $\frac{1}{25}$ Image	70.0%	100%	100%
Bottom Right $\frac{1}{50}$ Image	70.0%	90.0%	100%
Swap Bottom Right Mean	**76.0%**	**98.0%**	**100%**

Table E.2: General tampering detection sensitivity analysis - quarter-scale imagery.

Image Tampering	**Single Zone** (% Correct)	**16 Zones** (% Correct)	**100 Zones** (% Correct)
Authentic Control	**100%**	**100%**	**100%**
Center $\frac{1}{12}$ Image	70.0%	95.0%	100%
Center $\frac{1}{16}$ Image	70.0%	90.0%	100%
Center $\frac{1}{20}$ Image	65.0%	95.0%	100%
Center $\frac{1}{25}$ Image	65.0%	90.0%	100%
Center $\frac{1}{50}$ Image	65.0%	90.0%	100%
Swap Center Mean	**67.0%**	**92.0%**	**100%**
Top Left $\frac{1}{12}$ Image	70.0%	90.0%	100%
Top Left $\frac{1}{16}$ Image	70.0%	90.0%	100%
Top Left $\frac{1}{20}$ Image	70.0%	90.0%	100%
Top Left $\frac{1}{25}$ Image	70.0%	90.0%	100%
Top Left $\frac{1}{50}$ Image	70.0%	90.0%	100%
Swap Top Left Mean	**70.0%**	**90.0%**	**100%**
Top Right $\frac{1}{12}$ Image	70.0%	90.0%	100%
Top Right $\frac{1}{16}$ Image	70.0%	90.0%	100%
Top Right $\frac{1}{20}$ Image	70.0%	85.0%	100%
Top Right $\frac{1}{25}$ Image	70.0%	90.0%	100%
Top Right $\frac{1}{50}$ Image	70.0%	90.0%	100%
Swap Top Right Mean	**70.0%**	**89.0**	**100%**
Bottom Left $\frac{1}{12}$ Image	75.0%	90.0%	100%
Bottom Left $\frac{1}{16}$ Image	75.0%	95.0%	100%
Bottom Left $\frac{1}{20}$ Image	75.0%	90.0%	100%
Bottom Left $\frac{1}{25}$ Image	75.0%	85.0%	100%
Bottom Left $\frac{1}{50}$ Image	70.0%	85.0%	100%
Swap Bottom Left Mean	**74.0**	**89.0%**	**100%**
Bottom Right $\frac{1}{12}$ Image	70.0%	95.0%	100%
Bottom Right $\frac{1}{16}$ Image	70.0%	90.0%	100%
Bottom Right $\frac{1}{20}$ Image	75.0%	90.0%	100%
Bottom Right $\frac{1}{25}$ Image	75.0%	90.0%	100%
Bottom Right $\frac{1}{50}$ Image	65.0%	90.0%	100%
Swap Bottom Right Mean	**71.0%**	**91.0%**	**100%**

Table E.3: General tampering detection sensitivity analysis - sixteenth-scale imagery.

BIBLIOGRAPHY

[1] Google, "Biometrics — android open source project," 2021. [Online]. Available: https://source.android.com/security/biometric

[2] S. Liu, "Facial recognition global market size 2025," Apr 2021. [Online]. Available: https://www.statista.com/statistics/1153970/worldwide-facial-recognition-revenue/#statisticContainer

[3] P. Wasnik, K. B. Raja, R. Ramachandra, and C. Busch, "Assessing face image quality for smartphone based face recognition system," in *2017 5th International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2017, pp. 1–6.

[4] L. Pascu, "Biometric facial recognition hardware present in 90% of smartphones by 2024: Biometric update," Jan 2020. [Online]. Available: https://www.biometricupdate.com/202001/biometric-facial-recognition-hardware-present-in-90-of-smartphones-by-2024

[5] L. Omar and I. Ivrissimtzis, "Evaluating the resilience of face recognition systems against malicious attacks." 2015.

[6] A. B. Thabet and N. B. Amor, "Enhanced smart doorbell system based on face recognition," in *2015 16th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA)*. IEEE, 2015, pp. 373–377.

[7] M. Calvello, "22 eye-opening facial recognition statistics for 2020," Oct 2019. [Online]. Available: https://www.g2.com/articles/facial-recognition-statistics

[8] A. Opitz and A. Kriechbaum-Zabini, "Evaluation of face recognition technologies for identity verification in an egate based on operational data of an airport," in *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2015, pp. 1–5.

[9] Reservations.com, "Facial recognition statistics in airports: Survey shows 43% approve, 33% disapprove," Jan 2020. [Online]. Available: https://www.reservations.com/blog/resources/facial-recognition-airports-survey/

[10] J. Oliva, "Genesis gv60 will recognize your face to unlock the car," Sep 2021. [Online]. Available: https://www.motor1.com/news/533678/genesis-gv60-facial-recognition/

[11] A. Ng, "China tightens control with facial recognition, public shaming," Aug 2020. [Online]. Available: https://www.cnet.com/news/in-china-facial-recognition-public-shaming-and-control-go-hand-in-hand/

[12] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "Retinaface: Single-stage dense face localisation in the wild," *arXiv preprint arXiv:1905.00641*, 2019.

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

[14] E. Newton and S. Schuckers, "Recommendations for presentation attack detection (pad): Mitigation of threats due to spoof attacks." [Online]. Available: https://www.nist.gov/system/files/documents/2020/09/03/10_ibpc-prez-fido-ssanden-v5.pdf

[15] T. Soukupova and J. Cech, "Eye blink detection using facial landmarks," in *21st computer vision winter workshop, Rimske Toplice, Slovenia*, 2016.

[16] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting pulse from head motions in video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3430–3437.

[17] S. Liu, B. Yang, P. C. Yuen, and G. Zhao, "A 3d mask face anti-spoofing database with real world variations," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2016, pp. 100–106.

[18] A. Lagorio, M. Tistarelli, M. Cadoni, C. Fookes, and S. Sridharan, "Liveness detection based on 3d face shape analysis," in *2013 International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2013, pp. 1–4.

[19] H. Farid, "Image forgery detection," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 16–25, March 2009.

[20] S. P. Mohanty, N. Ranganathan, and R. K. Namballa, "Vlsi implementation of visible watermarking for secure digital still camera design," in *17th International Conference on VLSI Design. Proceedings.* IEEE, 2004, pp. 1063–1068.

[21] H. Wang, S. Z. Li, and Y. Wang, "Face recognition under varying lighting conditions using self quotient image," in *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.* IEEE, 2004, pp. 819–824.

[22] D. Deb, L. Best-Rowden, and A. K. Jain, "Face recognition performance under aging," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 46–54.

[23] S. Biswas, G. Aggarwal, N. Ramanathan, and R. Chellappa, "A non-generative approach for face recognition across aging," in *2008 IEEE Second International Conference on Biometrics: Theory, Applications and Systems*. IEEE, 2008, pp. 1–6.

[24] S. Devices, "Troubleshoot problems with windows hello on surface." [Online]. Available: https://support.microsoft.com/en-us/surface/troubleshoot-problems-with-windows-hello-on-surface-3346b3eb-6dd5-3d5e-e219-d47e790ef09b

[25] Y. Chen, Y. Xie, L. Song, F. Chen, and T. Tang, "A survey of accelerator architectures for deep neural networks," *Engineering*, vol. 6, no. 3, pp. 264–274, 2020.

[26] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," *arXiv preprint arXiv:1812.08685*, 2018.

[27] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 121–135, 2017.

[28] N. Sampat, S. Venkataraman, T. Yeh, and R. L. Kremens, "System implications of implementing auto-exposure on consumer digital cameras," in *Sensors, Cameras, and Applications for Digital Photography*, vol. 3650. International Society for Optics and Photonics, 1999, pp. 100–107.

[29] S.-H. Lam and C.-W. Kok, "Demosaic: Color filter array interpolation for digital cameras," in *Pacific-Rim Conference on Multimedia*. Springer, 2001, pp. 1084–1089.

[30] W. Yu, "An embedded camera lens distortion correction method for mobile computing applications," *IEEE Transactions on Consumer Electronics*, vol. 49, no. 4, pp. 894–901, 2003.

[31] G. L. Friedman, "The trustworthy digital camera: Restoring credibility to the photographic image," *IEEE Transactions on consumer electronics*, vol. 39, no. 4, pp. 905–910, 1993.

[32] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, vol. 1. Ieee, 2001, pp. I–I.

[33] O. Déniz, G. Bueno, J. Salido, and F. De la Torre, "Face recognition using histograms of oriented gradients," *Pattern recognition letters*, vol. 32, no. 12, pp. 1598–1603, 2011.

[34] H. Jiang and E. Learned-Miller, "Face detection with the faster r-cnn," in *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*. IEEE, 2017, pp. 650–657.

[35] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[36] S. Liao, X. Zhu, Z. Lei, L. Zhang, and S. Z. Li, "Learning multi-scale block local binary patterns for face recognition," in *International Conference on Biometrics*. Springer, 2007, pp. 828–837.

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

[38] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *2015 IEEE international conference on image processing (ICIP)*. IEEE, 2015, pp. 2636–2640.

[39] R. Shao, X. Lan, and P. C. Yuen, "Deep convolutional dynamic texture learning with adaptive channel-discriminability for 3d mask face anti-spoofing," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 748–755.

[40] D. F. Smith, A. Wiliem, and B. C. Lovell, "Face recognition on consumer devices: Reflections on replay attacks," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 736–745, 2015.

[41] D. K. Nilsson, U. E. Larson, and E. Jonsson, "Efficient in-vehicle delayed data authentication based on compound message authentication codes," in *2008 IEEE 68th Vehicular Technology Conference*. IEEE, 2008, pp. 1–5.

[42] V. M. Potdar, S. Han, and E. Chang, "A survey of digital image watermarking techniques," in *INDIN'05. 2005 3rd IEEE International Conference on Industrial Informatics, 2005.* IEEE, 2005, pp. 709–716.

[43] S. Cook, "Facetec: Why the samsung s10 proves liveness detection is needed in facial recognition." [Online]. Available: https://www.businesschief.com/technology-and-ai/facetec-why-samsung-s10-proves-liveness-detection-needed-facial-recognition

[44] Firebox, "Freak masks™ - stretchy personalised face masks." [Online]. Available: https://firebox.com/products/freak-masks-stretchy-personalised-face-masks

[45] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 2018, pp. 1–6.

[46] R. E. Overill, "Denial of service attacks: Threats and methodologies," *Journal of Financial Crime*, 1999.

[47] L. O'Gorman, "Comparing passwords, tokens, and biometrics for user authentication," *Proceedings of the IEEE*, vol. 91, no. 12, pp. 2021–2040, 2003.

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

[49] K. Patel, H. Han, and A. K. Jain, "Secure face unlock: Spoof detection on smartphones," *IEEE transactions on information forensics and security*, vol. 11, no. 10, pp. 2268–2283, 2016.

[50] N. Uchida, T. Shibahara, T. Aoki, H. Nakajima, and K. Kobayashi, "3d face recognition using passive stereo vision," in *IEEE International Conference on Image Processing 2005*, vol. 2. IEEE, 2005, pp. II–950.

[51] F. Tsalakanidou, F. Forster, S. Malassiotis, and M. G. Strintzis, "Real-time acquisition of depth and color images using structured light and its application to 3d face recognition," *Real-Time Imaging*, vol. 11, no. 5-6, pp. 358–369, 2005.

[52] N. N. Lakshminarayana, N. Narayan, N. Napp, S. Setlur, and V. Govindaraju, "A discriminative spatio-temporal mapping of face for liveness detection," in *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*. IEEE, 2017, pp. 1–7.

[53] K. Kollreider, H. Fronthaler, M. I. Faraj, and J. Bigun, "Real-time face detection and motion analysis with application in "liveness" assessment," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 548–558, 2007.

[54] G. Chetty and M. Wagner, "Audio-visual multimodal fusion for biometric person authentication and liveness verification," in *ACM International Conference Proceeding Series*, vol. 163, 2006, pp. 17–24.

[55] X. Li, J. Chen, G. Zhao, and M. Pietikainen, "Remote heart rate measurement from face videos under realistic situations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 4264–4271.

[56] S.-Y. Wang, S.-H. Yang, Y.-P. Chen, and J.-W. Huang, "Face liveness detection based on skin blood flow analysis," *symmetry*, vol. 9, no. 12, p. 305, 2017.

[57] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 1. IEEE, 2005, pp. 786–791.

[58] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*. IEEE, 2012, pp. 1–7.

[59] S. B. Nikam and S. Agarwal, "Gabor filter-based fingerprint anti-spoofing," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2008, pp. 1103–1114.

[60] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 746–761, 2015.

[61] R. Shao, X. Lan, and P. C. Yuen, "Joint discriminative learning of deep dynamic textures for 3d mask face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 4, pp. 923–938, 2018.

[62] A. Aggarwal and M. Kumar, "Image surface texture analysis and classification using deep learning," *Multimedia Tools and Applications*, vol. 80, no. 1, pp. 1289–1309, 2021.

[63] Z. Wang, Z. Yu, C. Zhao, X. Zhu, Y. Qin, Q. Zhou, F. Zhou, and Z. Lei, "Deep spatial gradient and temporal depth learning for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5042–5051.

[64] Z. Zhang, D. Yi, Z. Lei, and S. Z. Li, "Face liveness detection by learning multispectral reflectance distributions," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2011, pp. 436–441.

[65] Z. Akhtar and G. L. Foresti, "Face spoof attack recognition using discriminative image patches," *Journal of Electrical and Computer Engineering*, vol. 2016, 2016.

[66] A. Agarwal, D. Yadav, N. Kohli, R. Singh, M. Vatsa, and A. Noore, "Face presentation attack with latex masks in multispectral videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 81–89.

[67] H. Steiner, A. Kolb, and N. Jung, "Reliable face anti-spoofing using multispectral swir imaging," in *2016 international conference on biometrics (ICB)*. IEEE, 2016, pp. 1–8.

[68] A. Castiglione, G. Cattaneo, G. De Maio, A. De Santis, G. Costabile, and M. Epifani, "The forensic analysis of a false digital alibi," in *2012 Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*. IEEE, 2012, pp. 114–121.

[69] A. Anjos and S. Marcel, "Counter-measures to photo attacks in face recognition: a public database and a baseline," in *2011 international joint conference on Biometrics (IJCB)*. IEEE, 2011, pp. 1–7.

[70] B. Chesney and D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *Calif. L. Rev.*, vol. 107, p. 1753, 2019.

[71] reflectai, "reflect - first ever realistic face swap app," 2020. [Online]. Available: https://reflect.tech/

[72] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[73] M. Goljan, J. Fridrich, and T. Filler, "Large scale test of sensor fingerprint camera identification," in *Media Forensics and Security*, E. J. D. III, J. Dittmann, N. D. Memon, and P. W. Wong, Eds., vol. 7254, International Society for Optics and Photonics. SPIE, 2009, pp. 170 – 181. [Online]. Available: https://doi.org/10.1117/12.805701

[74] M. Goljan, M. Chen, P. Comesaña, and J. Fridrich, "Effect of compression on sensor-fingerprint based camera identification," *Electronic Imaging*, vol. 2016, no. 8, pp. 1–10, 2016.

[75] J. Lukas, J. Fridrich, and M. Goljan, "Digital camera identification from sensor pattern noise," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 205–214, June 2006.

[76] G. Chierchia, G. Poggi, C. Sansone, and L. Verdoliva, "Prnu-based forgery detection with regularity constraints and global optimization," in *2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP)*, Sep. 2013, pp. 236–241.

[77] P. Korus and J. Huang, "Multi-scale analysis strategies in prnu-based tampering localization," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 4, pp. 809–824, April 2017.

[78] B. Mahdian and S. Saic, "Using noise inconsistencies for blind image forensics," *Image and Vision Computing*, vol. 27, pp. 1497–1503, 09 2009.

[79] A. Levandoski and J. Lobo, "Image forgery detection: developing a holistic detection tool," 2020.

[80] D. Cozzolino and L. Verdoliva, "Camera-based image forgery localization using convolutional neural networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 1372–1376.

[81] ——, "Noiseprint: a cnn-based camera model fingerprint," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 144–159, 2019.

[82] A. Gunawan, H. Lovenia, and A. Pramudita, "Deteksi pemalsuan gambar dengan ela dan deep learning," 10 2018.

[83] L. Guarnera, O. Giudice, and S. Battiato, "Deepfake detection by analyzing convolutional traces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 666–667.

[84] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces (GUI)*, vol. 3, no. 1, 2019.

[85] S. Tariq, S. Lee, and S. S. Woo, "A convolutional lstm based residual network for deepfake video detection," *arXiv preprint arXiv:2009.07480*, 2020.

[86] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3677–3685.

[87] J. Naruniec, L. Helminger, C. Schroers, and R. M. Weber, "High-resolution neural face swapping for visual effects," in *Computer Graphics Forum*, vol. 39. Wiley Online Library, 2020, pp. 173–184.

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

[89] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," *arXiv preprint arXiv:1906.06876*, 2019.

[90] A. Chintha, A. Rao, S. Sohrawardi, K. Bhatt, M. Wright, and R. Ptucha, "Leveraging edges and optical flow on faces for deepfake detection," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2020, pp. 1–10.

[91] S. Zhang, X. Wang, A. Liu, C. Zhao, J. Wan, S. Escalera, H. Shi, Z. Wang, and S. Z. Li, "A dataset and benchmark for large-scale multi-modal face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 919–928.

[92] M. Asim, Z. Ming, and M. Y. Javed, "Cnn based spatio-temporal feature extraction for face anti-spoofing," in *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*. IEEE, 2017, pp. 234–238.

[93] M. Firbank, M. Oda, and D. T. Delpy, "An improved design for a stable and reproducible phantom material for use in near-infrared spectroscopy and imaging," *Physics in Medicine & Biology*, vol. 40, no. 5, p. 955, 1995.

[94] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, B. C. Van Esesn, A. A. S. Awwal, and V. K. Asari, "The history began from alexnet: A comprehensive survey on deep learning approaches," *arXiv preprint arXiv:1803.01164*, 2018.

[95] R. O. Green, "Retrieval of reflectance from calibrated radiance imagery measured by the airborne visible/infrared imaging spectrometer (aviris) for lithological mapping of clark mountains, california," in *Annual JPL Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) Workshop*, vol. 2, 1990, pp. 90–54.

[96] K. Man and I. Ashdown, "Accurate colorimetric feedback for rgb led clusters," in *Sixth International Conference on Solid State Lighting*, vol. 6337. International Society for Optics and Photonics, 2006, p. 633702.

[97] R. Basri and D. W. Jacobs, "Lambertian reflectance and linear subspaces," *IEEE transactions on pattern analysis and machine intelligence*, vol. 25, no. 2, pp. 218–233, 2003.

[98] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[99] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[100] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[101] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, and the scikit-image contributors, "scikit-image: image processing in Python," *PeerJ*, vol. 2, p. e453, 6 2014. [Online]. Available: https://doi.org/10.7717/peerj.453

[102] N. Ahmed, T. Natarajan, and K. R. Rao, "Discrete cosine transform," *IEEE transactions on Computers*, vol. 100, no. 1, pp. 90–93, 1974.

[103] M. J. Er, W. Chen, and S. Wu, "High-speed face recognition based on discrete cosine transform and rbf neural networks," *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 679–691, 2005.

[104] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[105] G. D. Simanjuntak, K. N. Ramadhani, and A. Arifianto, "Face spoofing detection using color distortion features and principal component analysis," in *2019 7th International Conference on Information and Communication Technology (ICoICT)*. IEEE, 2019, pp. 1–5.

[106] A. Rahimi and B. Recht, "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning," *Advances in neural information processing systems*, vol. 21, 2008.

[107] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *2017 international conference on engineering and technology (ICET)*. Ieee, 2017, pp. 1–6.

[108] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[109] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[110] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[111] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[112] FLIR, "Blackfly s usb3," 2022. [Online]. Available: https://www.flir.com/products/blackfly-s-usb3/?model=BFS-U3-51S5P-C&vertical=machine%2Bvision&segment=iis

[113] N. Efrat, D. Glasner, A. Apartsin, B. Nadler, and A. Levin, "Accurate blur models vs. image priors in single image super-resolution," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2832–2839.

[114] R. Widenhorn, M. M. Blouke, A. Weber, A. Rest, and E. Bodegom, "Temperature dependence of dark current in a ccd," in *Sensors and Camera Systems for Scientific, Industrial, and Digital Photography Applications III*, vol. 4669. SPIE, 2002, pp. 193–201.

[115] R. A. Boie and I. J. Cox, "An analysis of camera noise," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 14, no. 06, pp. 671–674, 1992.

[116] U. Shin, J. Park, G. Shim, F. Rameau, and I. S. Kweon, "Camera exposure control for robust robot vision with noise-aware image quality assessment," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1165–1172.

[117] B. Anusorn and C. Nopporn, "Light source estimation using feature points from specular highlights and cast shadows," *International Journal of Physical Sciences*, vol. 11, no. 13, pp. 168–177, 2016.

[118] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[119] T. Gloe and R. Böhme, "The dresden image database for benchmarking digital image forensics," *Journal of Digital Forensic Practice*, vol. 3, no. 2-4, pp. 150–159, 2010.

[120] J. Dong, W. Wang, and T. Tan, "Casia image tampering detection evaluation database," in *2013 IEEE China Summit and International Conference on Signal and Information Processing*, 2013, pp. 422–426.

[121] C. Meij and Z. Geradts, "Source camera identification using photo response non-uniformity on whatsapp," *Digital Investigation*, vol. 24, pp. 142–154, 2018.

[122] "Matlab vision toolbox," 2017, the MathWorks, Natick, MA, USA.

[123] X. Zhang and Y. Gao, "Face recognition across pose: A review," *Pattern recognition*, vol. 42, no. 11, pp. 2876–2896, 2009.

[124] E. Zhou, Z. Cao, and J. Sun, "Gridface: Face rectification via learning local homography transformations," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[125] L. Hu, M. Kan, S. Shan, X. Song, and X. Chen, "Ldf-net: Learning a displacement field network for face recognition across pose," in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*. IEEE, 2017, pp. 9–16.

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

[127] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[128] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[129] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks." in *ICML*, no. 3, 2016, p. 7.

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

[131] X. Yin and X. Liu, "Multi-task convolutional neural network for pose-invariant face recognition," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 964–975, 2017.

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

[133] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[134] Mut1ny, "Face/head segmentation dataset commercial purpose edition," Jul 2021. [Online]. Available: https://store.mut1ny.com/product/face-head-segmentation-dataset-pro?v=7516fd43adaa

[135] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[136] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.