Detection Methods against Digital Image Attacks for Secure Computer Vision

Sadu Chiranjeevi

Detection Methods against Digital Image Attacks for Secure Computer Vision

Thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

by

Sadu Chiranjeevi

Under the Supervision of

Prof. Pradip K. Das

Department of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
Guwahati 781039, India
June 2023

Dedicated to

My Family

For their blessings, constant inspiration and love

Declaration

I certify that

a. The work contained in this thesis is original and has been done by myself under the general supervision of my supervisor.

b. The work has not been submitted to any other institute for any degree or diploma.

c. Whenever I have used materials (data, theoretical analysis, results) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

d. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving the required details in the references.

Place: IIT Guwahati

Date:

Sadu Chiranjeevi
Research Scholar
Dept. of Computer Science and Engineering,
Indian Institute of Technology Guwahati,
Guwahati 781039, India.

CERTIFICATE

This is to certify that this thesis entitled " ***Detection Methods against Digital Image Attacks for Secure Computer Vision****" being submitted by* ***Sadu Chiranjeevi*** *to the Department of Computer Science and Engineering, Indian Institute of Technology Guwahati, is a record of bonafide research work under my supervision and is worthy of consideration for the award of the degree of Doctor of Philosophy of the Institute.*

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree or diploma.

Place: IIT Guwahati

Date:

Prof. Pradip K. Das
Department of Computer Science and Engineering,
Indian Institute of Technology Guwahati,
Guwahati 781039, India.

Acknowledgements

I would like to take this opportunity to thank many individuals who directly or indirectly supported me during my PhD process.

Foremost, I would like to express my deepest gratitude to my advisor, Prof. Pradip K. Das, for his constant support and encouragement to strive for excellence. Despite his preoccupation with several assignments, he has been kind enough to spare his valuable time and gave me the necessary council and guidance. His (i) ability to systematically disseminate scientific problems, (ii) attention to details, (iii) work ethics and (iv) promptness in responding to emails are some of the many virtues that I will always strive to instill in me. I am thankful for all that I learned from him and especially for accepting me to work with him when I needed.

I also want to thank my doctoral committee, Prof. Shivashankar B. Nair, Dr. Pinaki Mitra and Dr. D. Udaya Kumar, for their valuable comments and suggestions during my PhD.

I am also thankful to the faculty, Dr. R. Inkulu, Dr. Aryabhatta Sahu, Dr. Santosh Biswas, Dr. V. Vijaya Saradhi, Dr. Arajit Sur, Dr. Benny George and Prof. G. Sajith for imparting knowledge through various courses. A special thanks to Dr. Amit Awekar. Without his initial guidance on research basics and support, I would not have been writing this dissertation.

I also want to thank the department technical officers, technical superintendents and administrative staff for their wholehearted and unconditional support.

I am thankful to my host organization, Rajiv Gandhi University of Knowledge Technologies (RGUKT) Nuzvid, Andhra Pradesh for allowing me to continue my research work at IIT Guwahati and to all my colleagues for their continuous support. A special thanks to the higher-ups, RGUKT Nuzvid, for being so supportive always.

I would also like to express my heartfelt gratitude to the Director, the Deans and other management of IIT Guwahati, whose collective effort has made this institute a place for world-class studies and research. I am thankful to all the faculty and the staff of the Department of Computer Science and Engineering for the support received.

I owe a great deal to my parents for being a constant source of support and guidance. Their patience and sacrifice will remain an inspiration throughout my life. I am also grateful to my brothers, sisters-in-law, nephews and nieces for their love, constant inspiration and encouragement. They were always there during the ups and downs. Without their moral support, none of this would have been possible.

I want to thank my wife, Madhuri who has supported, encouraged and believed in me during this long journey that culminated in this dissertation. I would not have made it this far without her sacrifice and patience. A special note of thanks to my daughter Aradhya, who was the source of inspiration to work hard and enjoy my life during this journey, all of which helped me become more efficient at my work. They will always inspire me to move forward toward my destination.

I am beholden to my undergraduate students and collaborators Venkatesh, Sudha, Ratnam, Anusha, Noor Hasan, Gopi and Adithya, without whose constructive discussions, motivation and support, the work might not be possible in this duration.

Finally, I thank all my friends, Bala Prakasa Rao Killi, Bhagath, Deepak, Vamshali, Kiran Kumar, Siva Ram, Velu Chamy, Sandeep, Satya Prakash, Rakesh, Ghalib, Manoj, Hema, Sukarn, Saptharsi, Surajit, Rajesh, Panthadeep, Abhishek, Awnish, Sunil, Ashish, Jogen, Jagan, Pradeep Bhale and Pappu only to name a few, with whom I have spent most of the time.

Place: IIT Guwahati

Date: **Sadu Chiranjeevi**

Abstract

In today's digital age, our everyday life is filled with digital multimedia data as one of the primary forms for communication. Such data can be generated, processed, stored and transmitted in digital format in a very easy manner due to the widespread availability of inexpensive cameras, computers and user-friendly editing tools. As a result, Computer Vision (CV) systems supported by Machine Learning (ML) and Deep Learning (DL) techniques are now pervasive to process such multimedia content and have influenced every domain of life, ranging from security from various malware and attacks, healthcare and finance. However, with modern technologies in sophisticated editing tools and DL models, it becomes a critical task to protect CV systems from digital image attacks. *This thesis focuses on detecting a spectrum of digital attacks at the image level.*

Digital images play a pivotal role for carrying important information in many real-world fields. With developments in user-friendly editing tools and DL models, manipulating or attacking digital images becomes an easy task. These attacks have become advanced enough to trick CV systems and deceive Human Vision (HV) systems. Therefore, authentication of digital images is necessary. The thesis mainly focuses on (i) detection of face swap attacks (ii) detection of Copy-Move Forgery (CMF) attacks and (iii) detection of facial adversarial attacks.

Face swapping transfers the face of a source image to the face of a destination image or vice-versa while preserving photo realism. Although it has many applications, including computer gaming and entertainment, it could also be used for malicious or fraudulent purposes. We propose a method to create face swap attacks on original images and a technique to defend against them. Augmented 81-facial landmark points are extracted for creating the face swap attacks. The features are provided to Support Vector Machines (SVMs). The proposed detection method detects face swap attacks with 95% accuracy on a real-world dataset.

In CMF attacks, the attacker copies some regions of the image and pastes them into one or more regions of the same image. This attack's main aim is to cover or emphasize essential scenes in an image. We propose a detection method for such

forgery regions based on Binary Robust Invariant Scalable Keypoints (BRISK) and Speeded Up Robust Features (SURF) descriptors. Both fused features are matched and clustering is performed to reduce false positives. The proposed method is tested on real-world copy-move datasets. Experimental results show that our method is robust against various geometric transformations and precisely determines the forged regions.

ML models and especially DL models have impressively performed on perceptual tasks over the past few years. However, these models remain vulnerable to carefully crafted small perturbations, popularly known as adversarial attacks. Adversarial attacks modify an input by adding small perturbations to cause the classifier to misclassify the input. Such attacks become more problematic when they are used for pedestrians and in autonomous vehicles. Therefore, the detection of adversarial attacks is essential for the rightful and confident usage of DL-based solutions in the real world. As face provides a rich source of information, we propose novel defense methods to detect different types of adversarial facial attacks. The proposed defenses are evaluated on real-world datasets and experimental results show that they are robust against a wide range of adversarial face attacks.

Contents

List of Figures **16**

List of Tables **19**

List of Abbreviations **21**

1 Introduction **23**
1.1 Motivation of the Research Work 28
1.2 Contributions of the Thesis 29
1.2.1 Detection of Augmented Facial Landmarks-based Face Swapping 30
1.2.2 Detection of Copy-Move Forgery Attacks 30
1.2.3 Defense Methods against Facial Adversarial Attacks 31
1.2.4 Image Restoration for Improving Facial Adversarial Robustness 32
1.2.5 Organisation of the Thesis 32

2 Background and Literature Survey **34**
2.1 Computer Vision 34
2.1.1 Basic Structure of Computer Vision 35
2.1.2 Computer Vision Applications 37
2.1.3 Computer Vision Challenges 40
2.2 Digital Image Attacks 41
2.2.1 Entire Face Synthesis 42
2.2.2 Identity Swap 43
2.2.3 Face Morphing 44
2.2.4 Attribute Manipulation 45
2.2.5 Expression Swap 46
2.2.6 Copy-Move Forgery Attacks 46
2.2.7 Adversarial Attacks 48
2.3 Physical Attacks 50
2.3.1 Print Attacks 51
2.3.2 Replay Attacks 52
2.3.3 Disguise or Makeup Attacks 52

2.3.4 Mask Attacks . . . 53
2.4 Need for New Detection Methods . . . 53
2.5 Related Works . . . 56
2.5.1 Face Swap Attacks Detection . . . 56
2.5.2 Copy-Move Forgery Attacks Detection . . . 58
2.5.3 Defense Methods against Adversarial Attacks . . . 61
2.6 Summary . . . 65

3 Detection of Augmented Facial Landmarks-based Face Swapping 67
3.1 Problem Formulation . . . 68
3.2 Proposed Face Swapping . . . 69
3.2.1 Augmented Facial Landmarks Detection . . . 69
3.2.2 Face Swapping . . . 71
3.3 Proposed Face Swap Attack Detection . . . 75
3.4 Experimental Results . . . 76
3.4.1 Dataset . . . 76
3.4.2 Performance of the Proposed Face Swapping . . . 76
3.4.3 Performance of the Proposed FSAD . . . 79
3.5 Conclusions . . . 80

4 Detection of Copy-Move Forgery Attacks 82
4.1 SURF Features . . . 83
4.1.1 Integral Image . . . 83
4.1.2 Keypoints Detection . . . 84
4.1.3 Orientation Assignment . . . 85
4.1.4 Feature Descriptor Generation . . . 85
4.2 BRISK Features . . . 86
4.3 Keypoints Clustering . . . 87
4.4 Proposed CMFD Method . . . 87
4.5 Experimental Results . . . 88
4.5.1 Dataset . . . 88
4.5.2 Evaluation Metrics . . . 89
4.6 Conclusions . . . 92

5 Defense Methods against Facial Adversarial Attacks 94
5.1 Problem Definition . . . 96
5.2 Intensity-based Facial Adversarial Attacks . . . 96
5.2.1 Adversarial Attacks Generation . . . 97
5.2.2 Defense Method . . . 100
5.2.3 Experimental Results . . . 100
5.3 Geometry-based Facial Adversarial Attacks . . . 101

5.3.1 Adversarial Attacks Generation . . . 102
5.3.2 Error Level Analysis . . . 103
5.3.3 Defense Method . . . 104
5.3.4 Experimental Results . . . 104
5.4 Evaluating Robustness of Intensity-based and Geometric-based Adversarial Attacks . . . 106
5.5 Conclusions . . . 107

6 Image Restoration for Improving Facial Adversarial Robustness 109
6.1 Motivation . . . 109
6.2 Adversarial Attacks Generation . . . 111
6.3 Feature Denoising and Deep Image Restoration Networks . . . 111
6.4 Proposed Model for Improving Facial Adversarial Robustness . . . 112
6.4.1 Algorithm Description . . . 115
6.5 Experimental Results . . . 116
6.5.1 FFHQ Dataset . . . 116
6.5.2 Performance of the Proposed Defense Method . . . 118
6.6 Conclusions . . . 119

7 Conclusions and Future Directions 121
7.1 Future Directions . . . 123

List of Figures

1.1 Face attacks on facial images [1] 24
1.2 Examples for face swap attacks: Source image, destination image and face swap attack row-wise respectively 25
1.3 Copy-move forgery photograph released by Iran (appeared in The New York Times in July, 2008) [2] 26
1.4 Original images (in the first row) and their adversarial images 28

2.1 Major components of computer vision systems 35
2.2 Examples of entire face synthesis [3] 43
2.3 Examples for face swap attack: source image, destination image and face swap attack row-wise respectively 44
2.4 Example of a face morphing image in the middle of the first image and third image [4] 45
2.5 Examples of the attribute manipulation generated using FaceApp [3, 5] 46
2.6 Real and fake examples of expression swap [3]. Images are extracted from videos of FaceForensics++ database [6] 47
2.7 An example copy-move forgery attack and its detection [7] 47
2.8 Typical workflow for CMFD [8] 59

3.1 (a) The benchmark 68 facial landmarks by [9], (b) SFM to extract additional landmarks and (c) Augmented 81-facial landmarks 70
3.2 (a) A source image and (b) The augmented 81 facial landmark points 71
3.3 The pipeline of the proposed face swapping approach 71
3.4 (a) Convex hull of facial landmark points and (b) Triangles using Delaunay Triangulation 72
3.5 (a) A triangle of the source face, (b) A triangle of the destination face and (c) Warped triangle after affine transformation 72
3.6 (a) The destination image, (b) Face swap before blending and (c) Final face swap image 72
3.7 The procedure of the proposed FSAD approach 75
3.8 Source image, destination image and face swap image (from top to bottom in each column) 77

3.9 Source image, destination image, face swap with 68-facial landmarks and face swap with augmented 81-facial landmarks (from top to bottom in each column) 78
3.10 Source image, destination image, face swap with 68-facial landmarks and face swap with augmented 81-facial landmarks in row-wise 78
3.11 Incorrect face swaps: Source image, destination image and face swap image (from top to bottom in each column) 79

4.1 Integral image calculation by rectangular region 84
4.2 The BRISK sampling pattern with $N = 60$ 86
4.3 An overview of the proposed CMFD system; features extraction, clusters and detection results 88
4.4 Performance of the proposed method against various CMF attacks . . 90

5.1 The FGSM for adversarial image generation [10] 98
5.2 The procedure for proposed defense method 100
5.3 Original images (first row) and adversarial images generated with P-FGSM (second row) 101
5.4 Geometry-based adversarial face image generation by FLM [11] . . . 102
5.5 The overall procedure of the proposed defense method against geometry-based adversarial attacks 104
5.6 Some of the results for geometry-based face attacks generated using FLM. The first row represents original images from the CelebA dataset 105
5.7 Performance comparison of the proposed defense on all classification models 105

6.1 (a) Feature maps of clean and adversarial images and (b) Feature maps of adversarial images before and after denoising [12] 110
6.2 The overall procedure of the proposed defense method 113
6.3 The procedure for the extraction of WLMP features 113
6.4 Examples of original images, adversarial images generated using FLM, restored adversarial images by BL filter and restored adversarial images by BL+SR (from top to bottom in each column) 115
6.5 Examples of original images, adversarial images generated using P-FGSM, restored adversarial images by BL filter and restored adversarial images by BL+SR (from top to bottom in each column) . 116

List of Tables

3.1 Overall face swap attack detection results 80

4.1 Performance of the proposed CMFD in terms of TPR and FPR on varying threshold t 91
4.2 Performance comparison of the proposed CMFD 91

5.1 Performance of the proposed defense method 101
5.2 Overall results of the proposed defense method 105
5.3 Robustness comparison of intensity-based and geometric-based adversarial attacks on CelebA dataset 107

6.1 Overall results of the proposed defense method on CelebA Dataset with P-FGSM attacks 117
6.2 Overall results of the proposed defense method on FFHQ Dataset . . 118

List of Abbreviations

AFR	Automatic Face Recognition
API	Application Program Interface
APIs	Application Program Interfaces
AR	Augmented Reality
BL	Bilateral Filtering
BoW	Bag of Words
BRISK	Binary Robust Invariant Scalable Keypoints
CAT	Computerized Axial Tomography
CelebA	CelebFaces Attributes
CMF	Copy-Move Forgery
CMFs	Copy-Move Forgeries
CNN	Convolutional Neural Network
CNNs	Convolutional Neural Networks
CV	Computer Vision
CMFD	Copy-Move Forgery Detection
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DCT	Discrete Cosine Transform
DL	Deep Learning
DNN	Deep Neural Network
DNNs	Deep Neural Networks
DyWT	Dyadic Wavelet Transform
EDSR	Enhanced Deep Super-Resolution Network
ELA	Error Level Analysis
FGSM	Fast Gradient Sign Method
FLD	Facial Landmark Detection
FLM	Fast Landmark Manipulation
FMT	Fourier Mellin Transforms
FPR	False Positive Rate
FR	Face Recognition
FSAD	Face Swap Attacks Detection
GAN	Generative Adversarial Network

LIST OF ABBREVIATIONS

GANs	Generative Adversarial Networks
GFLM	Grouped Fast Landmark Manipulation
HAC	Hierarchical Agglomerative Clustering
HOG	Histogram of Oriented Gradients
HV	Human Vision
JSMA	Jacobian-based Saliency Map Attacks
k-NN	k-Nearest Neighbors
LBP	Local Binary Patterns
LBP-HF	Local Binary Patterns Histogram Fourier
LR	Logistic Regression
ML	Machine Learning
MRI	Magnetic Resonance Imaging
NN	Neural Network
NNs	Neural Networks
PA	Presentation Attack
PAs	Presentation Attacks
PAI	Presentation Attack Instrument
PAIs	Presentation Attack Instruments
PCA	Principal Component Analysis
PGD	Projected Gradient Descent
PSNR	Peak Signal-to-Noise Ratio
P-FGSM	Private Fast Gradient Sign Method
RF	Random Forest
RANSAC	RANdom SAmple Consensus
SFM	Surrey Face Model
SIFT	Scale Invariant Feature Transform
SOTA	State-Of-The-Art
SR	Image Super Resolution
SURF	Speeded Up Robust Features
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SVMs	Support Vector Machines
TPR	True Positive Rate
WLMP	Weighted Local Magnitude Patterns

Chapter 1

Introduction

In today's digital era, our day-to-day lives are filled with digital multimedia content as one of the primary forms of communication. Due to the availability of low-cost smartphones, cameras, computers and user-friendly editing tools, such content can be generated, processed, stored and transmitted in digital format in a very easy way. As a result, CV systems supported by ML and DL are now widespread to process such multimedia content for decision making. These systems have an impact on every domain of life, including security from various attacks and malware, industry, military, finance and healthcare. The enormous success of CV can be attributed to improvements in the algorithms, the accessibility of powerful computing resources, and the availability of massive datasets for various purposes. However, with modern technologies and DL algorithms, it becomes a challenging task to protect CV systems from digital image attacks. Thus, besides the technical and economic advantages, the rising of digital information has produced challenging problems with multimedia security and reliability.

Digital images play a vital role and are one of the most shared forms of digital multimedia content. They reveal information that is sensitive to a user such as their age, gender, dressing style, the presence of people and their relationship [13].

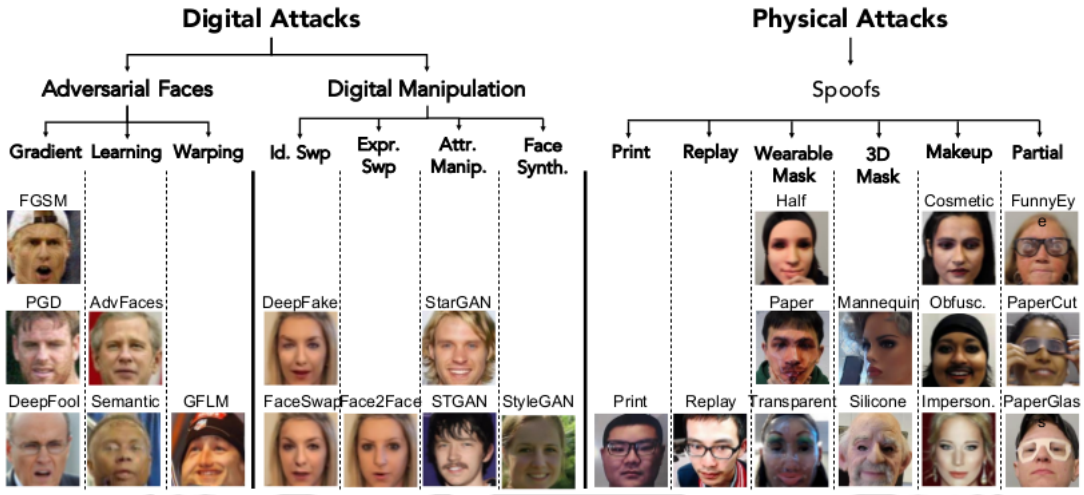


Figure 1.1: Face attacks on facial images [1]

These images are used in various applications such as surveillance, courts of law as a piece of evidence for crimes, journalism, scientific publications and medical imaging [2]. With the development of many image editing tools like Snapchat, Paint, etc. and DL models such as Generative Adversarial Networks (GANs) and Convolutional Neural Networks (CNNs), it becomes easy even for non-professionals to manipulate and create digital image attacks or fake images. A digital image attack modifies the content of the original digital image and creates a fake digital image. Its intention is to either cover or emphasize important regions of the image or create a digital attack for illegal purposes. For example, mobile applications and open software such as ZAO[1] and FaceApp[2] have opened a facility to anyone to generate digital attacks either in images or videos, without any expertise in the field. Such digital attacks could be used for malicious or fraudulent purposes such as targeted commercial, political advertising, viral Internet memes and pornography [14]. For example, DeepFakes [15] produces videos in which people perform actions or say things that never occurred. The most harmful applications of DeepFakes

[1]https://apps.apple.com/cn/app/id1465199127

[2]https://apps.apple.com/gb/app/faceapp-ai-face-editor/id1180884341

involve financial frauds, hoaxes, fake news and fake pornography [16]. Deb *et al.* [1] presented a diversity of attacks on face images as shown in Fig. 1.1.

Face swap is one of such digital attacks. It transfers the face from a source image to the face of a target image or vice-versa while preserving attributes such as facial expression, lighting, color, head pose, etc. Its intention is to create a face swap attack either in a video or image. Fig. 2.3 shows examples of face swap attacks. It has numerous applications in computer games, cinematic entertainment, face emotion recognition [17], preserving privacy [18] and entertainment [19]. However, it is also associated with viral Internet memes [14] and could also be used for malicious or illegal purposes. Zhang *et al.* [20] proposed the first work to address the detection of face swapping. They detected SURF [21] keypoints and performed k-means clustering over all selected SURF descriptors from the training data to form a Bag of Words (BoW) model. The features are provided into either linear or non-linear based ML models to predict their authenticity. It remains an active problem as face swap results appear more realistic and unedited.

Figure 1.2: Examples for face swap attacks: Source image, destination image and face swap attack row-wise respectively

One of the most actively investigated attacks in digital image forensics is the CMF attack. It copies some regions of an image and pastes them into one or more

parts of the same image and creates a forgery image. The purpose of this attack is to either hide or emphasize important objects of the original image and creates one more tampered image intentionally. It is a widely used attack to create tampered images because it is more complicated to detect as some features of the forged region such as noise and color are highly similar to the remaining regions of the image. Fig. 1.3 shows an example of CMF attack. The photo had appeared in Press July 2008, showing that four Iranian missiles were launched. Later, it was proved that three out of four were real and the other is fake. The original image is on the left side. On the right side, the forged image in which the original and forged regions are encircled. Fridrich *et al.* [22] proposed the first work based on Discrete Cosine Transform (DCT) for detection of Copy-Move Forgeries (CMFs). The image is partitioned into fixed-sized overlapping blocks at raster-scan and DCT is computed on each block. The feature vectors are lexicographically sorted to match and reduce the matching search space. Euclidean distance is used to compute the similarity between the feature vectors. Christlein *et al.* evaluated various Copy-Move Forgery Detection (CMFD) methods [23]. It is difficult to detect CMF attacks because the source and destination images for forgery are the same [24].

Figure 1.3: Copy-move forgery photograph released by Iran (appeared in The New York Times in July, 2008) [2]

In recent years, ML and DL models in particular have gained popularity in various applications including image classification [25], natural language processing [26], object detection [27], sentiment analysis [28] and multi-modal [29]. Despite their success in many applications, these models remain vulnerable to a particular class of malicious attacks known as adversarial attacks [10, 30]. Adversarial attacks are intentionally created by adding small perturbations to the input to cause the classifier for misclassification with high confidence. Such attacks can involve small but pathological modifications causing ML models to misclassify while humans can barely notice the perturbations. To determine if adversarial images are outside the distribution of the training data, Grosse *et al.* [31] employed a model-agnostic statistical test. They note that adversarial samples produced by some attacks can be located in various areas of the output surface as opposed to typical inputs and can be identified via statistical analysis. Detecting adversarial attacks is still a challenging task because DL-based adversarial attacks are intentionally created with key properties as follows: (i) maintaining the usefulness of images such that humans cannot identify the distortion; (ii) hiding the distortion such that a method cannot determine it; and (iii) preventing the deduction of a mapping between the true class and the class of the distorted image assigned by the classifier. That is, adversarial attacks are generated so that the impact of distortion on the adversarial image should be imperceptible, irreversible and undetectable.

A privacy violation may occur when a classifier infers sensitive information without user consent from fake digital images. Therefore, it is imperative to be able to protect digital content from digital image attacks to guarantee its security and truthfulness. The research community is active in this area, developing sophisticated and precise solutions for protection and authentication. In this thesis, we focus on detecting and mitigating the effect of digital image attacks to protect CV-based systems. We extracted encoded features from the input and provided them into

Figure 1.4: Original images (in the first row) and their adversarial images

either linear or non-linear ML models to detect swapped face images from the original. The fused features of SURF and BRISK descriptors are matched to address CMF attacks. As it is imperative to protect DL models from adversarial attacks, we propose various defense methods against a wide range of facial adversarial attacks.

1.1 Motivation of the Research Work

Given an image and the ease with which digital image attacks may be generated and distributed, it becomes increasingly difficult to know whether the image is real or if it is a fake one. Therefore, the verification of image originality is required in several CV applications such as scientific, military, media, entertainment, forensic, etc. [32]. Although there are many detection methods that have been proposed in the literature to defend against fake digital images from the original [3, 33], the research community is still active and working on advanced and accurate methods for protection and authentication for the following challenges:

- As face swap attack combines the attributes of both the source face and destination face images, the swapped faces appear realistic and look unedited. Detecting such attacks effectively is a challenging task and helps to protect CV systems such as Automatic Face Recognition (AFR), which is an imperative

feature for face authentication in smart devices, from illegal authentication.

- The attacker sometimes copies some of the regions of an image and pastes them onto one or more regions of the same image, popularly known as CMF attacks. As the attacker uses the same image as the source and destination for creating a CMF attack, properties like the color, noise and illumination conditions are expected to be well-matched between them. It is extremely difficult for the human eye to localize and detect such attacks. Therefore, there is a need to extract distinct features from images to accurately localize and detect such digital attacks.

- The face offers a rich amount of information, with just milliseconds of exposure being sufficient to draw implicit inferences about personal qualities like trustworthiness. The attacker intentionally adds small perturbations to an input face image to cause ML models for a wrong prediction with high confidence. Such attacks are known as adversarial attacks and due to their impact on ML and DL models, they pose a real threat to the real world. Therefore, there is a need to defend against such attacks to protect ML models as well as improve their adversarial robustness.

1.2 Contributions of the Thesis

Based on the motivation factors mentioned so far, we present a set of detection methods against digital image attacks. These methods detect whether an input image is real or fake. We briefly describe the problems addressed in this thesis. For each problem, we discuss the formulated detection method and mention the key observations from our evaluation. The details of these are presented in subsequent chapters of the thesis.

1.2.1 Detection of Augmented Facial Landmarks-based Face Swapping

We formulate the problem of detecting swapped face images from original images. We extract augmented 81-facial landmarks which include facial landmarks on the forehead as well. Full face swapping is performed based on the extracted augmented 81-facial landmarks of both the source face and destination face image. We extract encoded features Weighted Local Magnitude Patterns (WLMP) from images and feed them into different types of SVMs. We call the detection method as Face Swap Attack Detection (FSAD) for brevity. We evaluate the performance of our proposed FSAD on a real-world dataset and the key observations are summarized as follows:

- Linear-SVM and Polynomial-SVM achieve a precision of 96% and recall value of 94% for swapped face images with a detection accuracy of 95%.
- Both Sigmoid-SVM and Gaussian-SVM achieve the same values for precision and recall for both fake and original images with an accuracy of 74%.
- Linear-SVM and Polynomial-SVM outperform Sigmoid-SVM and Gaussian-SVM in terms of all parameter values.

1.2.2 Detection of Copy-Move Forgery Attacks

We formulate the problem of localization and detection of CMF attacks in digital images. We extract SURF and BRISK descriptors. We match both fused features and perform clustering using Hierarchical Agglomerative Clustering (HAC) to reduce false positives. The objective is to accurately localize and detect CMFs present in the image. We evaluate our detection method on real-world CMF datasets and experimental results are presented in terms of True Positive Rate (TPR) and False Positive Rate (FPR) with varying threshold t. We also compare the results

of our detection method with the State-Of-The-Art (SOTA) methods. The key observations from the numerical evaluation are as follows:

- Our detection method achieves the highest 98% TPR and lowest 7.5% FPR at the threshold 0.09.
- It outperforms some of the SOTA methods in terms of TPR and running time.

1.2.3 Defense Methods against Facial Adversarial Attacks

We formulate the problem of defending against facial adversarial attacks from clean images. We generate different kinds of facial adversarial attacks and also present the results obtained after facial adversarial attacks. We extract distinct feature descriptors from face images and provide them to various types of classification models to defend against facial adversarial attacks from clean images. We evaluate our defense methods on real-world datasets. The performance of our methods is demonstrated with different types of classifiers. The key observations are as follows:

- Linear SVM outperforms the remaining classifiers in terms of all evaluating parameters for detecting intensity-based facial adversarial attacks.
- WLMP features effectively highlight intensity-based adversarial noises in face images.
- Error Level Analysis (ELA) is effective in highlighting geometry-based adversarial noises in face images.
- Logistic Regression (LR) outperforms the remaining classification models in terms of all metrics with 0.99 precision, 1.00 recall, 1.00 F1-score and 99.75% accuracy for detecting geometry-based facial adversarial attacks.

1.2.4 Image Restoration for Improving Facial Adversarial Robustness

We formulate the problem of defending against a wide range of facial adversarial attacks and improving the facial adversarial robustness of classifiers. We generate facial adversarial attacks based on different kinds of adversarial methods. We restore the facial adversarial images using image restoration techniques. That is, we bring back images into the original space from the adversarial space by applying bilateral (BL) filtering and Super Resolution (SR). WLMP features are extracted and fed into various classifiers. We evaluate our defense method on real-world datasets. We also present experimental results before and after image restoration techniques. The performance of our method is demonstrated on different kinds of classifiers. The key observations from the numerical evaluation are as follows:

- Linear SVM shows its effectiveness in detecting facial adversarial images from the original with the highest accuracy of 98.75%.
- After employing image restoration such as BL followed by SR (BL+SR) to the adversarial images, the classification accuracy improves from 98.75% to 99% for Linear SVM.
- It is also observed that sometimes BL alone is sufficient to enhance the visual quality of images, which brings back the low-resolution adversarial images into the high-resolution original space.

1.2.5 Organisation of the Thesis

The rest of the thesis is organized as follows:

- In the next chapter, we present the background required to understand the problems we addressed. We also present the SOTA literature on the detection

of face swap attacks, CMF attacks and adversarial facial attacks.

- In Chapter 3, we address the problem of detecting face swap attacks from the original. The proposed method extracts WLMP features and provides them into a classifier to effectively detect swapped face images.
- In Chapter 4, we address the problem of detecting CMF attacks with various geometric transformations. The fused features of SURF and BRISK are used to accurately detect various CMF attacks.
- In Chapter 5, we propose defense methods against a wide range of facial adversarial attacks. The distinct feature analysis is used to effectively detect both intensity-based and geometry-based facial adversarial attacks from clean images.
- In Chapter 6, we propose a defense method with improved facial adversarial robustness. Image restoration techniques are used to improve the facial adversarial robustness of various classifiers.
- Finally, the thesis ends with a summary and future work in Chapter 7.

Chapter 2

Background and Literature Survey

This chapter briefly discusses the basic structure of CV, its applications and challenges in today's digital world. It also discusses various digital image attacks, their impact on CV systems and the need for new detection methods against such attacks. Finally, it discusses the SOTA methods related to detection methods against digital image attacks.

2.1 Computer Vision

In the current digital age, Internet users have surpassed half the world's population. One minute on the internet might not seem like a lot of time, but with billions of users using it every day, the statistics of what gets done are surprising [1]. For instance, in a minute, users share more than 2,40,000 images on Facebook and post approximately 70,000 images on Instagram. It aggregates to more than 1.8 billion image uploads per day [2]. It shows the impact of digital media content in everyone's life, irrespective of their domain of work. Moreover, digital media content has emerged as one of the main forms of communication and it is also straightforward to create, process, store

[1] https://www.stackscale.com/blog/internet-one-minute/

[2] https://www.kleinerperkins.com/perspectives/internet-trends-report-2018/

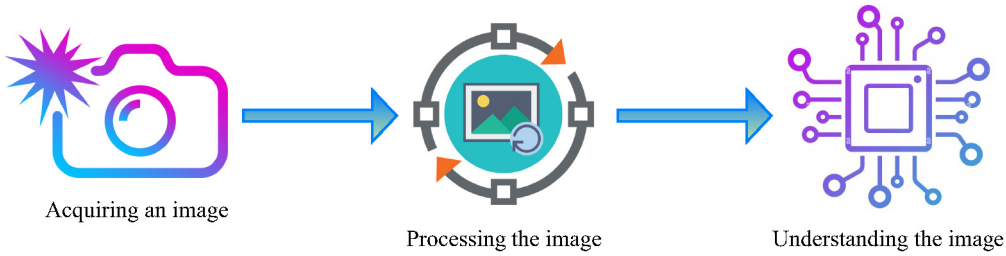


Figure 2.1: Major components of computer vision systems

and transmit it in digital format due to the wide variety of smartphones, modern editing tools, low-cost cameras and computers. Thus, CV systems, mainly supported by ML and DL models, are now widespread across all the domains to process such largely generated multimedia content for decision making.

2.1.1 Basic Structure of Computer Vision

"Computer vision is a field of computer science that trains computers to interpret and understand the visual world in the same way that human does". CV focuses on building digital systems that can process, analyze and make sense of visual data such as images and videos for decision making. The major components of the CV system are shown in Fig. 2.1.

Acquiring an Image: A digital image is produced by one or more image sensors such as radar, range sensors, tomography devices, ultrasonic cameras, etc., in addition to various kinds of light-sensitive cameras. Depending on the type of sensor used, the output image data can be a simple 2D image, a 3D amount, or an image series. As the internet developed in the 1990s, generating massive collections of images that were accessible online for analysis became easy. In particular, smartphone technologies with built-in cameras have filled up the globe with low to high-resolution images and videos. These expanding datasets made it feasible for computers to recognize specific individuals in images and videos.

Processing the Image: The digital image is processed to extract a specific piece of information such as texture, pixels, motion, etc. Based on the application, a choice is made regarding which regions of the image or image points are important for processing is taken at some point in the processing. For instance, the regions of image are as follows:

- Selection of a particular set of points of interest.
- Segmentation of one or more regions of the image which contain a particular object of interest.
- Application-specific parameter estimation, such as object pose or size.
- Image recognition-classification of a detected object into various groups.
- Image registration of two different views of the same object are compared and merged.

High-end hardware designed for computing is widely available. As a result, computing power becomes more accessible and more affordable. In particular, DL models such as CNNs are used to automate and process large collections of images easily by taking advantage of the software and hardware capabilities. These models are often trained by being initially provided with massive datasets with thousands of labeled or pre-identified images.

Understanding the Image: It is the interpretative step where an object is identified or classified. During the processing stage, the computer does intricate computations and formulates relationships with the components of the image to understand what it represents. The computer does this by using three levels of data: low level, intermediate level and high level. The low level comprises image primitives like regions, borders, or texture components. The intermediate level includes surfaces, boundaries and volumes. The high level includes scenes,

objects, or events. There are various types of CV systems that are used for various applications:

- **Image segmentation** divides a digital image into several subgroups known as image segments, which serves to simplify further processing or analysis of the image by decreasing the complexity of the original image.
- **Object detection** identifies a particular object in an image. Advanced object detection identifies many objects in an image. It uses (x, y) coordinates to form a bounding box around the objects.
- **Facial recognition** Facial recognition is a special type of object detection. It identifies or confirms the identification of individuals using their faces. It can be utilized to detect individuals in images or videos in real-time.
- **Edge detection** is a method used to detect the boundaries of an object to better interpret what is in the image. It detects edges by identifying discontinuities in brightness.
- **Pattern detection** is used to recognize repeated colors, shapes and other visual components in images.
- **Image classification** is a fundamental task that classifies images into different groups by assigning them a particular label.
- **Feature matching** is a type of pattern detection algorithm. It recognizes features of the same components across images with different viewpoints to classify them.

2.1.2 Computer Vision Applications

CV systems are now pervasive and have applications in a wide variety of fields that depend on computers to interpret images. It is the backbone of an autonomous

future across many industry sectors, including transportation, healthcare, agriculture, retail, manufacturing and more. Recently, Tesla announced it would transmit entirely to Tesla Vision[3], a camera-based autopilot system, retiring radar. Major applications of CV include:

- **Military** uses CV to enable a crucial technology for modern armies that help security systems identify enemy troops and improves guided missile systems' targeting capabilities. CV systems also provide battlefield intelligence used for tactical decision-making, military principles such as situational awareness heavily rely on image sensors. It is also having a key role in developing autonomous vehicles to traverse difficult terrain and recognize adversaries.

- **Healthcare** diagnostics relies heavily on the study of images, scans and photographs. CV helps to detect anomalies in imagery extracted from Computerized Axial Tomography (CAT) and Magnetic Resonance Imaging (MRI) scans with far higher precision than medical practitioners can get. It also helps to simplify the analysis of various medical images to prevent false diagnoses and reduce treatment costs.

- **Manufacturing industry** depends on CV systems for automatic inspection of faulty goods on the production line and remote inspection of pipelines and machinery. It helps to flag unusual events or discrepancies and optimize organizational and control processes. It also provides technology for predictive maintenance, package inspection, bar code scanning, monitoring and product assembly.

- **Education** uses CV for applications such as school logistic support, knowledge acquisition, attendance monitoring and regular assessments. CV-enabled webcams are used to monitor students during examinations and make unfair

[3]https://electrek.co/2021/05/25/tesla-vision-without-radar-warns-limitations-first/

practices easier to spot through the analysis of eye movements and body behavior. Initiations have started to enhance the perception of the learners through the use of CV, especially technologies such as Augmented Reality (AR) have grown due to online or remote education. "AR is an interactive experience of a real-world environment where the objects that reside in the real world are enhanced by computer-generated perceptual information, sometimes across multiple sensory modalities". Integrating AR assists students with various learning skills and enhances the efficacy of the classroom environment.

- **Agriculture** uses CV-based solutions for weeding, detecting plant health and advanced weather analysis. It has numerous applications, including drone-based crop monitoring, automatic spraying of pesticides, yield tracking and smart crop sorting and classification. These vision-based applications scan the crops' shape, color and texture for further analysis. Modern vision-based technologies enable farmers to cultivate ever-larger fields efficiently. If they are not properly monitored, plant diseases can lead to painful harvest losses and crop failures. CV also helps to analyze data generated using drones, satellites and remote sensors to estimate various parameters and monitor automatically.

- **Automotive industry** are developing autonomous cars for attending to various aspects of the real-world. Although the human driver is not yet replaced, autonomous vehicle technology has made significant progress over the past few years. The future of self-driving cars heavily relies on CV, especially DL models, to capture the imagination of the public. CV enables autonomous vehicles to interpret their surroundings such as road edges, traffic signs, objects, people, other vehicles, etc. Cameras record video from various perspectives and then provide it to CV software to process and make sense of the visual data in real-time.

- **Retail industry** such as Amazon and Flipkart use their digital platform's analysis capabilities to analyze customer behavior in detail and to optimize the user experience. CV systems help the retail industries from marketing and sales to customer service and retention. It also provides useful insights into consumer behavior and helps to up-sell and cross-sell.

- **Transportation** uses CV systems to detect traffic signal violators and allow law enforcement agencies to minimize unsafe on-road behavior. Intelligent sensing and processing solutions are also being used to detect speeding and wrong side driving violations, among other disruptive behaviors. It is also being used by intelligent transportation systems for traffic flow analysis.

- **Security industry** is a noteworthy driver for face detection solutions for detecting and preventing criminal activities. Detecting and recognizing faces in public is a contentious application of CV that is already being implemented in certain jurisdictions and banned in others. Although facial recognition is already in use at the personal level such as through smartphone applications, CV-based face recognition solutions are useful in tracking specific persons for security missions.

2.1.3 Computer Vision Challenges

Although CV has created a significant impact across all fields, it still faces challenges for two major reasons. First, the HV system is a complex and powerful system. It is difficult to replicate it using technology. Second, threats from the digital information revolution and modern technologies. The major challenges that CV systems face from the latter are as follows:

Fake Content: It becomes an easy task even for non-experts to create fake

content either in images, videos or text due to the availability of powerful user-friendly editing tools, low-cost cameras and the use of DL technologies. With such fake content, CV in the wrong hands can result in dangerous problems like many other contemporary technologies.

Adversarial Attacks: It modifies the content of an image by adding a small perturbation to it. It potentially misleads the ML model for misclassification. When an attacker creates such a faulty ML model, it is very difficult to identify and may seriously harm any system in the real world.

Reasoning Issue: Modern DL-based algorithms are complicated systems in which the functions are often unclear. When this happens, it is difficult to understand any task's logic, which makes it difficult for CV professionals to define any parameter in an image or video.

Privacy and Ethics: Globally, CV-powered monitoring poses a severe threat to people's privacy. It puts people at risk of unlawful data usage. Because of these issues, face recognition and detection are forbidden in several nations.

2.2 Digital Image Attacks

Digital images are one of the most shared forms of digital multimedia content. In earlier days, the amount and realism of digital image changes were constrained due to a lack of advanced editing tools, the need for specialized knowledge and the laborious and time-consuming procedure involved. For instance, the initial work [34] modified the lip movements of a speaker using a different audio track by drawing links between the subject's face structure and the sounds of the audio track. Nowadays, with advancements in many image editing tools like Snapchat, Paint, etc. and DL models

such as CNNs and GANs, it becomes easy even for non-professionals to manipulate and create digital image attacks or fake images. Digital attacks are modifications made to the original image that can cause a classifier to provide a different output from the original image. Recently, digital attacks on images and videos generated, in particular by DL-based [15] approaches, have become a great public concern[4]. In this section, we briefly review various digital image attacks.

2.2.1 Entire Face Synthesis

It generates whole non-existent facial images, often using a powerful Generative Adversarial Network (GAN). These methods achieve exceptionally realistic results, producing high-quality facial images for the observer. A GAN typically comprises two distinct Neural Networks (NNs) competing against one another in a minimax game: Generator G captures the distribution of the data and generates new samples, whereas Discriminator D calculates the probability that a sample will really originate from the training data (real) rather than G (fake). In order to produce high-quality fake samples, the training approach for G aims to increase the probability that D will make a mistake. D is discarded following the training process, while G is employed to produce fake content. The entire face synthesis has used this idea in recent years, increasing the realism of the modifications. Examples of entire face synthesis produced by StyleGAN [35] are shown in Fig. 2.2. The video gaming and 3D modeling businesses stand to gain from this manipulation, but it may also be used for undesirable purposes like developing incredibly convincing fake profiles on social media to spread false information.

[4]https://www.bbc.com/news/technology-49961089

Figure 2.2: Examples of entire face synthesis [3]

2.2.2 Identity Swap

It replaces the face of one subject (source) either in an image or video with the face of another subject (target) or vice-versa. The goal of identity swap is to produce realistic fake images or videos. Some examples of visual images for face swap are shown in Fig. 2.3. Although different sectors benefit from this kind of attack, especially in the film industry[5], it could also be used for illegal purposes such as financial frauds, hoaxes and the creation of celebrity pornographic videos, among many other illegal uses. The identity swap generation process typically involves the following stages for each frame of the source video: (i) face detection and cropping, (ii) extraction of intermediate representations, (iii) synthesis of a new face based on some driving signal (e.g., another face) and (iv) blending the generated face of the target subject into the source video. For identity swap manipulations, two distinct strategies are often taken into consideration:

- Traditional computer graphics-based approaches such as face swap.
- Novel DL-based algorithms known as DeepFakes.

[5]https://www.youtube.com/c/Shamook/featured

In face swap, face alignment, optimization and blending are used to swap the source subject's face with the target subject's face. In DeepFake [15], two autoencoders with a common encoder that has been trained to recreate training images of the source and target faces, respectively. The images are aligned and cropped using a face detector. The trained encoder and decoder of the source face are applied to the target face to produce a fake image. The output of the autoencoder is subsequently merged with the remainder of the image.

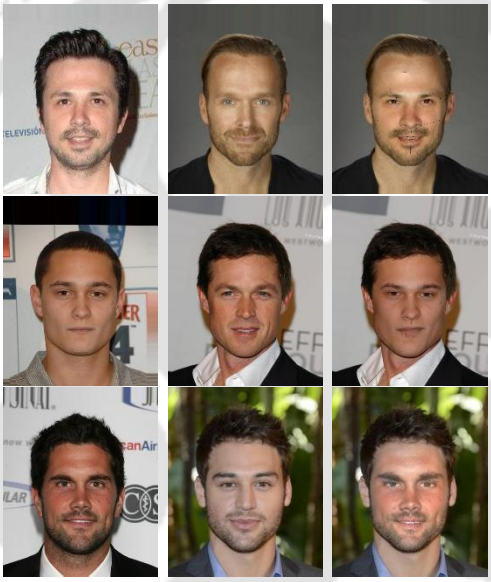


Figure 2.3: Examples for face swap attack: source image, destination image and face swap attack row-wise respectively

2.2.3 Face Morphing

Face morphing is a sort of digital face alteration that can be used to produce fake images that mimic the biometric data of two or more people [4]. The morphed face images are successfully verified against face images of individuals whose Face Recognition (FR) systems would be seriously threatened by them [36]. Examples of face morphing are shown in Fig. 2.4. In general, the creation of face morphing images

involves the following three steps in order: (i) finding the correlation between the faces of the various subjects. This is often done by first extracting facial landmark points such as the eyes, mouth, nose tips, etc.; (ii) then distorting the original face images of the subjects until the corresponding landmarks of the images are geometrically aligned; and (iii) performs blending to merge the color values of the warped images.

Figure 2.4: Example of a face morphing image in the middle of the first image and third image [4]

2.2.4 Attribute Manipulation

It is often referred to as face retouching or face editing. It modifies certain attributes of the face, such as skin, hair color, age, gender, adding spectacles, etc. [37]. Typically, this modification process is done using GAN such as the StarGAN [35]. The well-known mobile application for creating this kind of manipulation is FaceApp. With the use of this technology, users can virtually try on a wide variety of things, including glasses, cosmetics and hairstyles. Some examples of the attribute manipulation generated by FaceApp are shown in Fig. 2.5.

Figure 2.5: Examples of the attribute manipulation generated using FaceApp [3, 5]

2.2.5 Expression Swap

This manipulation, commonly referred to as face reenactment, involves modifying the subject's facial expression. Fig. 2.6 displays some visual examples taken from the FaceForensics++ database [6]. It could be leveraged for serious implications, e.g., a well-known video of Mark Zuckerberg making statements he never said[6].

2.2.6 Copy-Move Forgery Attacks

CMF is a widely used attack in digital images. It pastes some regions of the image to one or more parts of the same image and creates a forgery image. The purpose of this attack is to hide an important content of the original image and create one more forged digital image intentionally [38]. It is the most used attack because it is more complicated to detect as some features of the forged regions, such as noise and color are highly similar to the remaining regions of the image. Moreover, the source and destination images for forgery are the same [24]. An example CMF attack and its detection is shown in Fig. 2.7.

[6]https://www.bbc.com/news/technology-48607673

Figure 2.6: Real and fake examples of expression swap [3]. Images are extracted from videos of FaceForensics++ database [6]

Figure 2.7: An example copy-move forgery attack and its detection [7]

The creation of a CMF attack is simple. As the source and the target regions are both parts of the same image, characteristics like lighting, color temperature and noise are anticipated to be well-matched between the forged regions and the original. This attack could be used for targeted commercial and non-commercial advertisements, political advertisements, etc.

2.2.7 Adversarial Attacks

An adversarial image is a sample of an input image that has been very slightly altered with the intention of misclassifying it by an ML classifier. In several instances, a human observer may not even detect the modification, yet the classifier still fails to classify the adversarial image correctly. Security issues arise from adversarial attacks because they may be used to attack ML models even when the adversary does not have access to the underlying model. The approaches used to generate adversarial examples are broadly classified into three categories depending on the adversary's knowledge about the target classifier: 1) White-box adversarial attacks, 2) Black-box adversarial attacks and 3) Semi white-box adversarial attacks.

White-box Attacks: In white-box attacks, an attacker is assumed to have complete knowledge about the kind of neural network (NN) along with the number of layers of the target classification model. The attacker has access to the training data distribution and knowledge of the technique used for training, such as gradient-descent optimization. He has full information about the fully trained model architecture's parameters. The attacker uses the available information to determine the feature space where the model can be attacked or for which the model has a high error rate. Then the model is exploited by altering an input image using the adversarial example crafting method. The access to internal model weights for a white-box attack corresponds to a very strong adversarial attack.

Black-box Attacks: In black-box attacks, the information about the target classification model is not available to attackers. Attackers can only give input for the models and query their outcomes. For instance, when an adversarial image is given to the model, a label or a confidence score relating to another class of image is returned depending on the target classifier. The attackers usually observe such type of relationships between the input and output of the model. They can then use its flaws to attack the models for generating adversarial attacks.

Semi White-box Attacks: In these attacks, first a generative model is trained for creating adversarial samples in a white-box manner. Once it is trained, the adversary can use the trained model to craft adversarial samples in a black-box way. The authors trained GAN [39] to target the model of interest. Then, adversarial samples are crafted directly from the trained generative model.

Adversarial Goals: An adversary makes an effort to provide a classification system with an input adversarial image x^{adv} that leads to incorrect output. The incorrectness of the model leads to the adversary's goal being deduced. The following broad categories can be used to categorize adversarial goals based on their influence on the classifier's output integrity:

1. **Confidence Reduction:** The attacker attempts to lower the confidence score of the prediction of the target classifier for the given input. For example, an input image of a class label X can be predicted with less confidence having a low probability of the same class label X.

2. **Misclassification:** The adversary attempts to change an input example's output classification to a different class from the original class. For instance, any other class other than the class label X will be predicted for a real image of the class label X.

3. **Source/Target Misclassification:** The adversary tries to cause a certain target class to be the result of classification for specific input. For instance, an input image of class label X will be predicted as the class label Y by the classification model.

2.3 Physical Attacks

In the information age, automatic user access to services has gained more significance. As a result, ML-based biometric or FR systems have been developed to recognize people automatically in a variety of applications. However, such systems are sensitive to external attacks due to advanced technology, which could compromise their integrity [40]. External attacks on biometrics or FR systems are broadly categorized into: direct attacks and indirect attacks. Direct attacks are also referred to as physical attacks or spoofing attacks. Physical attacks are methods where a face's physical characteristics are altered before an image is taken. Some of the most important methods for physical attacks are presentation attacks (PAs), variances brought on by disguise or makeup and intentional plastic surgery. Direct attacks take place at the sensor stage when fake facial artifacts are presented. Such attacks include attacks on the system without the attacker being aware of the system's functionality, feature extraction techniques, or matching algorithms. These attacks are also known as sensor attacks. In indirect attacks, the attacker has to be aware of system knowledge to carry out attacks. These attacks affect feature extraction, matching, decision modules and database. For example, indirect attacks on the biometric systems need information about the internal working of the model, such as an attack on the communication system, feature extraction module, matching module, etc. They are generally treated as black-box attacks as the internal operation of the models is not openly known. Therefore, we present only different types of physical attacks.

PAs [41] can be used either to impersonate or to obfuscate a person in FR systems. Impersonation is an attack in which the adversary attempts to authenticate himself or herself as another user. It is used to access FR systems by copying a real user's facial attributes. Attacks related to this type of attack are prints, replays and masks. Obfuscation is used to conceal the identity of the user using a variety

of techniques, including the use of glasses, cosmetics, a masked face and facial hair. Attacks related to this type of attack generally involve different forms of disguises, such as wigs, glasses, tattoos and makeup. The tools used for PAs, including photos, videos and masks, are referred to as presentation attack instruments (PAIs). A general FR system recognizes authorized people in relation to the reference database and recognizes faces from the image or video input. The images produced by the sensor are altered and distorted in different ways by PAs. A spoof image may have different noise content than an actual image. It likely occurs distortions such as color distortion, surface reflection and share deformation. PAs have replicated face features such as masks, photographs or films to assist the adversary in breaching the security model if the FR lacks a detection module to distinguish between real and fake faces.

To avoid potential attacks and maximize their advantages for the users, it is crucial to comprehend the risks to which they are exposed and to analyze their vulnerabilities. PAs are broadly categorized into 2D and 3D attacks. Print and replay attacks are included in 2D attacks, whereas mask attacks are in 3D attacks.

2.3.1 Print Attacks

It is a most crucial type of attack in the PAs is a print or photo attack. Attackers are supposed to lack access to the recognition system's internals and gain entry by only projecting printed images of the targeted identity onto the input camera. There are several ways to obtain high-quality facial photos of the individuals who will be impersonated, including hidden cameras, social networks and online profiles. These images can then be presented to the sensor of the FR systems by being printed or shown on a screen. Print attacks include eye-cut photos, flat printed photos, warped photos and digital display of photos. These attacks are incredibly simple for a variety of reasons [42]. On the one hand, producing color photos of a real user's

face is inexpensive and simple. Alternatively, the images can be seen on a device's high-resolution screen, such as a smartphone, tablet, laptop, etc. On the other hand, the recent advance in social media platforms like Twitter, Facebook and Instagram makes it extremely simple to collect examples of real faces. Additionally, with the recent price and size reductions experienced by digital cameras, it is now possible to capture high-quality images of an authorized user by only utilizing a hidden camera. Even though such print attacks might appear to be too easy to work, many SOTA FR systems are susceptible to them [43].

2.3.2 Replay Attacks

The attacker gets a video of the real user they want to impersonate, plays it on any device such as a smartphone, tablet, laptop, etc., to reproduce the video and then displays it to the camera or sensor of the FR systems [44]. Such attacks are known as replay or video attacks. Due to the expansion of video sharing platforms (e.g. YouTube), social networks and even the usage of hidden cameras, it is extremely simple to get face videos of the users, just like in the case of print attacks. Replay attacks are harder to identify than photo attacks because they mimic not only the shape and texture of the face but also its dynamics, such as blinking eyes, facial movements and mouth. It is fair to infer that systems that are vulnerable to print attacks would perform even worse when subjected to replay attacks due to their increased level of complexity and that being resilient against photo attacks does not equate to being similarly powerful when subjected to replay attacks.

2.3.3 Disguise or Makeup Attacks

It is one kind of direct attack. Accessories such as hats, sunglasses and scarves can be used to impersonate or obfuscate either intentionally or unintentionally. Due to their strong similarity to the genuine face, makeup attacks are more difficult to

detect [45]. The three different makeup techniques used during data collecting are "Heavy Contour", "Pattern" and "Transformation". The first two techniques are used in three levels of intensity and are made to alter the normal shadows and the shape of the contours of the face. In the last technique, the participant's face is altered to mimic another identity, typically a well-known character.

2.3.4 Mask Attacks

The Presentation Attack Instrument (PAI) for this type of attack is a 3D mask of the user's face. The attacker displays the sensor/camera with a 3D reconstruction of the user's face. Mask attacks demand greater technical proficiency than earlier attacks, as well as access to more knowledge in order to create a convincing mask of the real person [46]. The simplest approach involves printing a 2D image of the user's face, which is then adhered to a deformable structure. A plastic bag or a t-shirt are two examples of this kind of construction. The attacker can then show the bag to the biometric sensor by placing it on his face. This technique can imitate some deformable facial patterns, making it possible to fool some basic 3D FR software. Face masks are more realistic in terms of texture, color and geometry than conventional 2D PAs. Different materials are used to create 3D masks. For example, solid or hard masks can be manufactured from plaster, paper, resin, or plastic, while soft masks are typically made of latex or silicon. Masks made of silicon or latex are flexible, soft and adapt to various facial sizes and shapes. They closely resemble the texture and color of real facial features.

2.4 Need for New Detection Methods

Despite the effectiveness of detecting fake images or videos has made great strides, there are a number of issues with the existing detection techniques that call

for caution. For example, CV-based systems are essential in several real-world applications. However, their security flaws related to adversarial samples might be utilized to control and compromise their application. Thus, some of key challenges demand new detection methods, including:

- **Evaluation of proposed attacks or defenses is not straightforward:** It is simple to assess traditional ML by calculating the loss on the test set, supposing that a training set and test set have been generated. Defenders facing adversarial attacks have an open-ended challenge where the attacker will supply inputs from an unidentified distribution. It is insufficient to benchmark a defense method against a single attack or even a group of attacks. Even if the defense method succeeds in such an experiment, it can still lose against a new attack that operates in a manner the defender did not foresee. A defense method should ideally be demonstrably solid, but ML in general and DL in specific are challenging to examine conceptually.

- **Emergence of modern technologies:** Digital image attack is incredibly simple and a frequent practice due to the availability of low-cost and open-source image handling tools such as Paint, Photoshop, Photoscape, PhotoPlus, GIMP, Pixelmator, etc. It has also become extremely difficult to detect visually whether a given image is the original or a modified version. Particularly DL-generated attacks could be extremely hard to identify with the naked eye. As a result, it is now simple for individuals and small organizations to create digital attacks and spread them widely in a short period of time, endangering the credibility of the news and the public's faith in social media.

- **Social media laundering:** Social media platforms such as Facebook, Twitter or Instagram are among the primary online networks utilized to disseminate digital multimedia information to the general public. Such content

is frequently modified before uploading to conserve network traffic or protect the privacy of the user. These alterations, which are typically referred to as social media laundering, eventually lead to an increase in false positive detection rates by removing hints regarding underlying forgeries. The majority of fake detection techniques use keypoints at the signal level, which are more vulnerable to social media laundering. In order to improve the effectiveness of false identification techniques for social media laundering, simulations of these effects should be carefully incorporated into training data. Evaluation databases should also be expanded to include information on social media laundered visual material.

- **Quality of DeepFake datasets:** DL-powered technologies make the availability of large-scale datasets which facilitates the creation of detection methods.

- **The rise of cloud services:** Many cloud service providers, including Google, AWS, Baidu, Alibaba, Azure, etc., offer DL Applicant Program Interfaces (APIs) for their clients to complete CV tasks without the need to train models and own a large amount of data. This is because DL frequently needs large training data and prolonged training times. Users of cloud services can use these APIs to verify photos for both profitable and non-profit purposes. For instance, Alibaba Cloud[7] and Azure[8] offer APIs to determine whether the images are legal or illegal (e.g., pornographic, violent). Such cloud service planning must be done in a setting with greater security.

- **Increasing demand for Big data applications:** The rising acceptance of big data applications which need for processing and sharing enormous amounts

[7]https://www.alibabacloud.com
[8]https://azure.microsoft.com

of data. This must be done in an increasingly secure manner to preserve the confidentiality of the data and project the processing models.

- **Model scalability** Another major concern is the inability of the current fake detection techniques to scale for large-scale platforms like social media [47]. In a real-world setting, inference time plays a crucial role in detection. The model is unlikely to be extensively employed in practical applications even if it is designed with great precision but a relatively long inference time. Therefore, there is a need for detection methods that can detect large amounts of fraudulent information in real-time and with high accuracy.

2.5 Related Works

In this section, we briefly discuss the survey of related works to detect digital image attacks from genuine images particularly face swap attacks, CMF attacks and adversarial attacks.

2.5.1 Face Swap Attacks Detection

There are many existing approaches for targeting face manipulation and its detection. The detection approaches that target image forgery may or may not work for face swapping. Thus, we review only the related works that address face swap attacks and their detection. Blanz *et al.* [48] proposed an approach that evaluates a 3D face model and its scene parameters, such as the focal length of the camera, the 3D orientation, position, etc., to properly exchange faces in images. The approach is similar to the morphable model in that it optimizes every model parameter while going from 3D to image. Bitouk *et al.* proposed a fully automatic face swap algorithm in images without the usage of 3D reconstruction methods [49]. It detects all faces that are present, aligns them to the coordinate system and selects

the target images from the large face library that are similar to the source image in pose and appearance. Then, the target images are modified to match the appearance of the source images in terms of color, lighting and postures. Mahajan *et al.* [50] demonstrated an algorithm that picks front-facing faces automatically and swaps them out with stock faces. In [51], the authors used Convolutional Neural Network (CNN) to capture the appearance of the target image from the unstructured image datasets. In this method, the network has to be trained for each target image and thus, it is not a practical solution for many applications. Chen *et al.* [52] proposed a method for replacing faces in referenced images that share similar traits and shapes with the input face. By adapting the reference face and its coordinating background to the input face, a triangulation-based technique is employed to distort the image. Numerous novel face swapping approaches have also been developed as a result of DL's practical success in image processing. Korshunova *et al.* [51] treated face swap as a process requiring style transfer. They viewed identity as the style and position and facial expression as the content. For image transformation, a CNN with several scale branches that operate on various image resolutions is employed. GAN-based methods have been proposed in [53–55], which produced impressive face swapping results.

Zhang *et al.* proposed an automated face swapping method and utilized SURF and BoW features rather than raw pixels for detecting swapped face images [20]. The performance is demonstrated with different types of classifiers like SVMs, Random Forest (RF) and Neural Networks (NNs) to discriminate the swapped face images from the genuine ones with an accuracy of 92%. The face swap quality is not evaluated against other datasets. Furthermore, the authors used a dataset consisting of only $10,000$ images which is relatively small compared to other works. Agarwal *et al.* proposed a novel feature descriptor, called WLMP, which is similar to Local Binary Patterns (LBP) and fed them into the Support Vector Machine

(SVM) classifier to detect face swap images [56]. Instead of images, they chose to target videos. They also produced a dataset of their own. Korshunov *et al.* proposed DL-based method for detecting swapped face images in videos [57]. They tested several DeepFakes detection techniques. Khodabakhsh *et al.* assessed the generalization potential of CNN-based and texture-based fake face detection systems [58]. For the evaluation, they employed a new dataset with $53,000$ images collected from 150 videos. They produced the swapped faces in their dataset using several methods. Smoothing and blending were utilized to appear the results of face swap more realistic. Rossler *et al.* evaluated various detectors in different scenarios [6]. Ding *et al.* proposed a DL-based model that uses transfer learning for the detection of swapped face images [59]. They also provided a large dataset containing $4,20,053$ images taken from 86 celebrity images.

2.5.2 Copy-Move Forgery Attacks Detection

When an image has been tampered with, its statistical characteristics will be changed. Then, the original image's statistical features are more distinct from the forged one. To detect tampered regions, the features of regions of the image are calculated and subsequently similarity checking and then matching the forged regions. Various block generation and keypoints generation based approaches have been developed over the years for addressing CMF attacks in digital images [60]. The majority of CMFD techniques adhere to the same essential steps [8] as displayed in workflow Fig. 2.8. The input image is subjected to the pre-processing process in the first stage. It enhances the picture data and features for more detection. The input image is converted into gray-scale and additional preparation, such as filtering or image resizing, can be optimized. After the preprocessing step, the feature extraction process to extract the picture's features is optimised. This feature extraction can be done with either block-based or key points-based techniques. Once the features are

extracted, an important process is to match identical features for marking forged regions. Then, the filtering process removes the forged matched features and finally detects if the image is attacked or not.

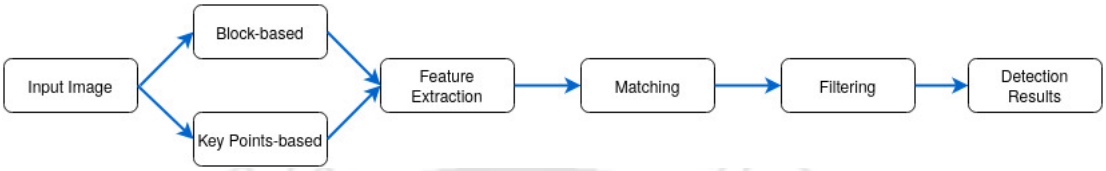


Figure 2.8: Typical workflow for CMFD [8]

In block-based methods, the tampered image is sliced into either overlapping or non-overlapping blocks of fixed size. Thereafter, feature vectors are computed from each block using different approaches. Extracted feature vectors are lexicographically sorted; therefore, matched feature vectors come close to each other. These feature vectors are then matched to each other to find the forgery regions. The authors of [61] first proposed a DCT-based method for the detection of forgeries. The image is partitioned into fixed-sized overlapping blocks at raster-scan and DCT is computed on each block. The feature vectors are lexicographically sorted to matching and reduce the matching search space. Euclidean distance is used to compute the similarity between the feature vectors. Popescu and Farid [62] leveraged Principal Component Analysis (PCA) to reduce the dimensional size of each feature vector to 32. This approach is robust against compression and noise addition. In [63], Fourier Mellin Transforms (FMT) based method is proposed for the detection of forgeries with feature vectors of size 45. The authors used a bloom filter instead of lexicographical sorting, which improves the forgery detection time. These features are rotation invariant only for some degree. In [64], a blur moment invariant features based method is proposed to detect the CMF in the presence of blur degradation, additive zero-mean noise or arbitrary contrast changes. In [65], the authors proposed a rotation invariant method based on Hu moments. The Gaussian pyramid is

applied to reduce the dimension of the forgery image. Then, Hu moment is used to extract features on the low-frequency image. Eigenvectors are then lexicographically sorted and matching is performed between them. In [66], the authors utilize rotation invariant Zernike moments. It is robust and gives significant performance against geometric transformations such as JPEG compression, additive noise and blurring. In [67], a robust method is proposed for blind CMFD based on an undecimated Dyadic Wavelet Transform (DyWT). In [7], the authors presented a method based on DCT and Singular Value Decomposition (SVD) to precisely locate the forgery regions in the images. A new method, PatchMatch, to detect and localize forgeries in images is proposed in [68]. They used nearest-neighbor search to efficiently deal with features and to achieve high robustness against scaling and rotations, especially over dense field images. The authors in [69] proposed a blind forgery detection method based on Local Binary Pattern Histogram Fourier Features (LBP-HF). They divided the forgery image into overlapping blocks and the features were extracted from each block.

In keypoints-based methods, keypoints are detected from the image and for each keypoint, feature descriptors are extracted. Thereafter, feature descriptors are matched for forgery region detection. Huang *et al.* [70] suggested an effective approach for detecting CMF based on Scale Invariant Feature Transform (SIFT) [71] features in images. Euclidean distance is used between the feature descriptors for matching the keypoints. These features are robust against various image post-processing transformations, such as rotation, scaling, additive noise and translation. Therefore, it is used in applications of various fields. In [72], a new method is proposed which is based on SIFT features [71] to detect duplicated and distorted regions in a digital image. Its performance is good on various image post-processing due to the robustness of SIFT feature descriptors. In [73], the authors presented a SIFT-based method for detecting multiple forgeries in a digital image. Additionally,

they used RANdom SAmple Consensus (RANSAC) to cluster the keypoints and select a set of inliers that are appropriate for a homography transform between the two clusters. Amerini *et al.* [2] utilized the SIFT [71] to detect the multiple CMFs. They performed clustering technique on spatial locations of matched points to make clusters and detect multiple forgery regions. In [74], the authors leveraged SURF [21] features for detecting tampered regions. It detects the keypoints and extracts 64 feature descriptors per keypoint and so it increases the speed. Another local features based method for the detection of forgeries is presented in [75]. In [2], the authors proposed an approach which is based on SIFT descriptors and performed HAC on spatial locations of matched points to make clusters and detect multiple forgery regions in the images. Additionally, they addressed CMFD for image slicing and also used RANSAC to identify the transformation. In [76], the authors utilized SIFT features and density-based spatial clustering of applications with noise (DBSCAN) clustering technique for detecting multiple forgery regions. Another contribution based on SIFT features is presented in [77] for the detection of CMFs.

2.5.3 Defense Methods against Adversarial Attacks

Adding a small and targeted perturbation on the input image converts the clean image into an adversarial image that can mislead a trained classification model for a false prediction confidently. Although DL models outperformed traditional classification models in a number of key areas, such as training big datasets and using strong computing resources, Szegedy *et al.* [30] demonstrated that these DL models are susceptible to these types of adversarial attacks.

Intensity-based Adversarial Attacks: These attacks especially aim to modify the intensity of an input image. Biggio *et al.* [78] proposed initial works for a simple gradient-based attack to systematically assess the security of traditional ML classification models such as SVM and NN. The authors demonstrated such attacks

with the MNIST dataset. Szegedy *et al.* [30] proposed a method for creating the first adversarial samples to attack Deep Neural Networks (DNNs), which is based on a box-constrained L-BFGS [79]. Goodfellow *et al.* [10] proposed a single-step approach for generating adversarial attacks, named as Fast Gradient Sign Method (FGSM). It is a fast and efficient attack showing that DL models are sensitive to such attacks. The authors used the gradient of classification loss related to the input image to add perturbation to the intensity of the original sample. To explore the sensitivity of ML models, many extensions to create intensity-based adversarial attacks have been developed so far. Kurakin *et al.* [80] proposed the first basic iterative approach to the FGSM [10]. Dong *et al.* [81] proposed a momentum-based iterative method to strengthen adversarial attacks. Rozsa *et al.* [82] used the actual value of the gradient rather than the sign of the gradient to improve the robustness of attacks against defenses. Papernot *et al.* [83] suggested an approach that creates Jacobian-based Saliency Map Attacks (JSMA) using the Jacobian matrix of the predicted classes for the input image. To reduce the number of pixels that must be changed during the attack, it builds a saliency map of the input's most valued pixels. Moosavi *et al.* [84] considered the decision boundary of a classification model around a particular data point x. The classifier gives a different prediction for x based on its path.

Further, a black-box attack to target Deep Neural Network (DNN) classifiers is proposed in [85]. Chen *et al.* [86] proposed an optimization-based approach for generating adversarial attacks where the adversary has no information about the target classifier. Ilyas *et al.* [87] presented an approach for creating black-box attacks, which is based on the gradient information from the outcomes of the classification model. A genetic approach for creating adversarial samples is proposed in [88]. A GAN-based semi white-box attack model is proposed in [89]. Deng *et al.* [90] proposed a method, called ArcFace, based on additive angular margin loss to determine highly distinctive features to maximize the classification of facial images.

Deb *et al.* [91] proposed a GAN-based approach, called Advfaces to create adversarial face images. The authors added small perturbations to the key facial features. In [92], the authors proposed a novel approach, SemanticAdv, to generate adversarial samples based on attribute-conditioned image editing.

In general, high-frequency elements are added to the input samples in almost all intensity-based adversarial attacks and the amount of distortion is controlled by $l_p - norm$ similarity measure. However, the similarity metric $l_p - norm$ is not a suitable metric and it fails to ensure that all adversarial images fall in the same space as the original images. As a result, it increases the sensitivity of such attacks, particularly where the attacker has no restrictions on time for evaluating the authenticity of the input images.

Geometry-based Adversarial Attacks: Xiao *et al.* [93] proposed a method based on spatial transformation instead of $l_p - norm$ of intensity values to generate adversarial attacks. They defined a flow or displacement field f per pixel to create the adversarial image. Dabouei *et al.* [11] proposed a novel and fast method to generate geometrically-perturbed faces. The authors used facial landmarks to generate such adversarial attacks.

Adversarial Attacks Detection: Since the emergence of adversarial examples, several methods have been proposed to mitigate the effect and detect such attacks. One of the most common methods to protect classifiers from a wrong prediction is to detect adversarial images from original images. These methods can not directly predict an input label of the model. Instead, such methods first determine whether the input sample is original or fake. The classifier then can not predict the input class label if the input is adversarial. Thus, these defense methods effectively discriminate adversarial images and mitigate the effect of adversarial attacks

on classification models from a wrong prediction.

Goodfellow *et al.* [10] proposed an approach to create adversarial samples based on FGSM and utilize them for training the classifier to counteract such samples. Hendrycks *et al.* [94] proposed a statistical defense approach based on PCA. Madry *et al.* [95] investigated the adversarial robustness of NNs. They trained networks against a wide range of projected gradient descent (PGD) adversaries as reliable first-order adversaries. Tramer *et al.* [96] proposed a method to aggregate adversarial training to detect adversarial attacks. In [31], a complementary approach is introduced to identify adversarial inputs. The authors specially augmented the ML classifier with an extra output in which the classifier is trained to target all adversarial inputs. Gong *et al.* [97] proposed a simple binary classifier and trained it to detect adversarial inputs from clean data with an accuracy of over 99%. Metzen *et al.* [98] proposed to augment DNN with a small subnetwork for detection. The subnetwork is trained on a binary classifier to distinguish adversarial inputs from genuine data. This method mainly focused on making the binary classifier itself more robust to adversarial inputs. Massoli *et al.* [99] augmented the DL model with k-nearest neighbors (k-NN) to separate adversarial images. Agarwal *et al.* [100] proposed an approach based on pixel values and PCA as features and provided to the SVM classifier for the detection of image-agnostic adversarial perturbations. An improved approach for adversarial robustness based on feature de-noising blocks in networks is proposed in [12]. In [101], the authors proposed a defense method based on image SR. They performed SR on adversarial inputs to bring them back into the natural space. Recently, the authors proposed to use Private Fast Gradient Sign Method (P-FGSM) attacks and their performance is evaluated on different types of SVMs to detect adversarial face images from the clean images.

2.6 Summary

In summary, we discussed the basic structure of CV systems and their components. We also discussed applications and their challenges in the real-world. We discussed different types of digital and physical image attacks and their impact on CV systems. Although there are many methods to address digital image attacks, we further discussed why we need new detection methods. Then, we presented related works on face swap attacks, CMF attacks and adversarial attacks for their generation and detection.

It is observed that with modern technologies, creating digital fakes or attacks which are strong enough to deceive CV and HV systems becomes an easy task. These digital fakes pose a real threat in the real world. Therefore, detecting such fakes or attacks in digital images remains a challenging task and an active problem.

Chapter 3

Detection of Augmented Facial Landmarks-based Face Swapping

Face swapping or face replacement is important in many situations, such as the provision of privacy, video compositing, appearance transformation in portraiture and other artistic applications. Depending on the application, this problem's precise formulation changes with certain objectives. Formally, it replaces the face of the source with the face of the target while preserving the attributes of the source or vice versa. Due to its realistic and unedited results, it could also be used for illegal purposes such as financial frauds, hoaxes and the creation of celebrity pornographic videos, among many other illegal uses [57]. Thus, face swap attacks pose a real threat to CV-based FR and biometric systems which provide essential features for security in many modern devices.

This chapter describes the detection of face swap attacks. It first extracts augmented 81-facial landmark points covering the landmark points on the forehead and performs face swapping. Given an image, the first 68 facial landmarks are extracted using the technique proposed by Kazemi and Sullivan [9] and additional 13 facial landmark points that cover the forehead are extracted based on Surrey

Face Model (SFM) [102] as shown in Fig. 3.1. Next, face swapping is performed based on the extracted augmented 81-facial landmark points. Finally, Face Swap Attack Detection (FSAD) is proposed to detect the swapped face images from the original images. It is based on WLMP features [56] and SVM [103]. The proposed system is evaluated on a real-world dataset. The experimental results show that it effectively performs face swapping and accurately does the detection.

The major contributions of the chapter are:

- We propose an approach to extract augmented 81-facial landmark points of a human face by using [9] and SFM [102].
- Face swapping is performed based on the extracted augmented facial landmarks of source and destination image faces.
- FSAD is proposed to detect whether the image has undergone a face swap attack or not.
- Finally, the performance of the proposed system is demonstrated by different types of SVMs and experimental results are presented.

The remainder of the chapter is organized as follows: Section 3.1 presents the problem formulation. The proposed methods for face swapping and its detection are explained in Sections 3.2 and 3.3, respectively. The dataset used is presented in Section 3.4.1. The results of the proposed methods are elaborated in Sections 3.4.2 and 3.4.3, respectively. We conclude the chapter in Section 3.5.

3.1 Problem Formulation

We take into account the situation where, given a single input facial image of any person T, we would want to replace that person's face with that of another person

S while preserving the original pose, gaze direction, facial expression, lighting and hairstyle of the input image T. The input image T is treated as the target or destination and S is as the source image. Given an input facial image I and let $C(.)$ be a classifier. Then, a detection method aims to detect the face swap attacks from the original, i.e., C(I) is either real or fake.

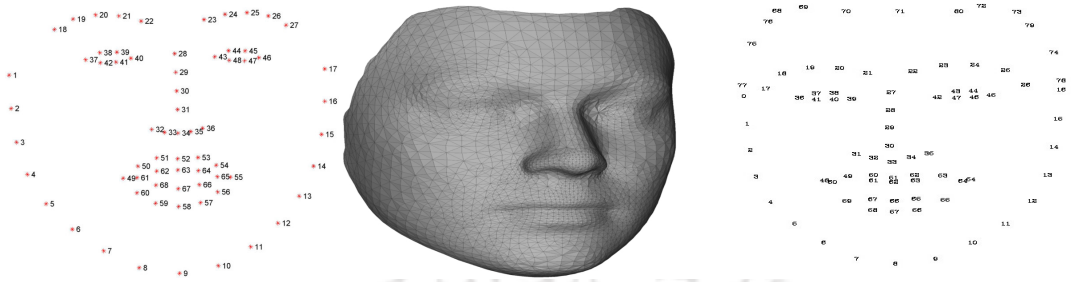
3.2 Proposed Face Swapping

Face swapping is challenging when swapping faces in unconstrained and arbitrarily selected images. In such cases, there is no guarantee of the similar appearance of expressions, viewpoints, genders or any attribute when faces are swapped.

The proposed method differs from the existing methods. It selects two random images and swaps the face of the source image with the face of the destination image instead of randomly searching for the destination image in the image library. It extracts augmented 81-facial landmark points that cover facial landmark points on the forehead instead of the 68-facial landmark points. The performance of the proposed detection approach is demonstrated with different types of SVM classifiers.

3.2.1 Augmented Facial Landmarks Detection

Facial landmarks are utilized to represent and localize the salient components of a face, such as eyes, eyebrows, nose, mouth and jawline. Facial Landmark Detection (FLD) identifies points of interest in an image of a human face. Detecting precise facial landmarks improves the performance of many CV and computer graphics applications such as face recognition, face animation, facial emotion detection, assessing gaze direction, facial emotion detection, augmenting face with graphics, face swapping, facial image synthesis, etc. [104–106]. For example, several face recognition approaches depend on the spatial locations of facial landmarks to align



(a) (b) (c)

Figure 3.1: (a) The benchmark 68 facial landmarks by [9], (b) SFM to extract additional landmarks and (c) Augmented 81-facial landmarks

faces from one image to another image. In this case, the detection of imprecise facial landmarks could lead to bad alignment and degrade the performance of face recognition. In other applications, 2D facial landmarks are utilized to deform 3D face meshes for realistic face performances. Some significant image-based facial landmarks detection methods are [107–109].

First, the faces of the source and destination images are detected using the Histogram of Oriented Gradients (HOG) and linear SVM [110]. Next, the (x, y)-coordinates of augmented 81-facial landmark points of the source and destination images are extracted as shown in Fig. 3.1c. The 68-facial landmark points are localized using a fast and accurate approach, which is based on an ensemble of regression trees [9] as shown in Fig. 3.1a and additional 13 facial landmark points that include the facial landmark points on the forehead are extracted by utilizing SFM [102] as shown in Fig. 3.1b. The extracted augmented 81-facial landmarks for a face image are shown in Fig. 3.2.

(a) (b)

Figure 3.2: (a) A source image and (b) The augmented 81 facial landmark points

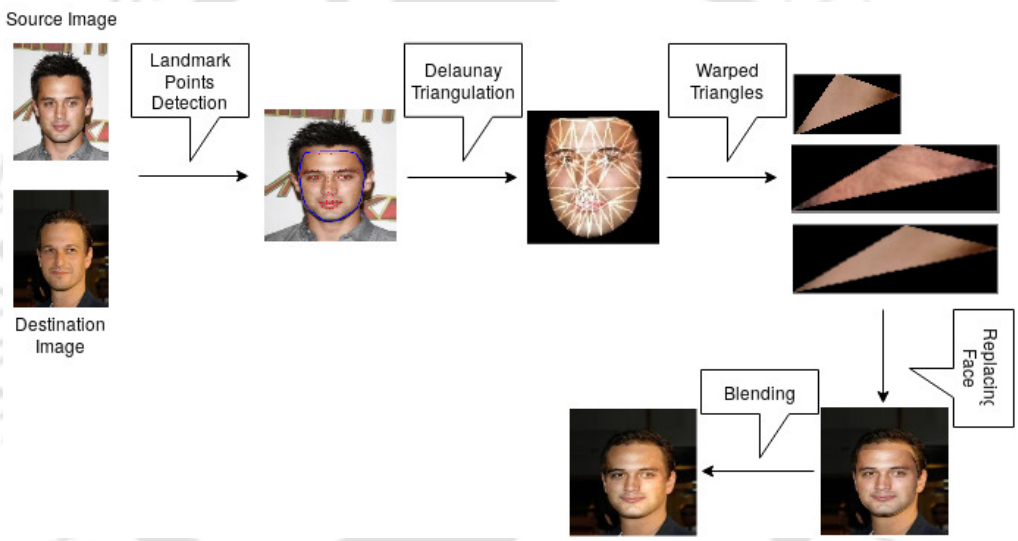
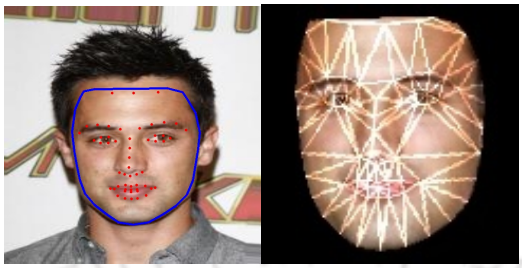


Figure 3.3: The pipeline of the proposed face swapping approach

3.2.2 Face Swapping

The pipeline of the proposed approach for face swapping is shown in Fig. 3.3. In this subsection, the complete process of how face swapping is performed between two face images is described in the following steps:

1. Once the facial landmark points are localized, the 3D faces can be approximated by considering the 2D planes of the facial landmark points. That is, a small area of the 2D plane can be transformed into another 2D plane that

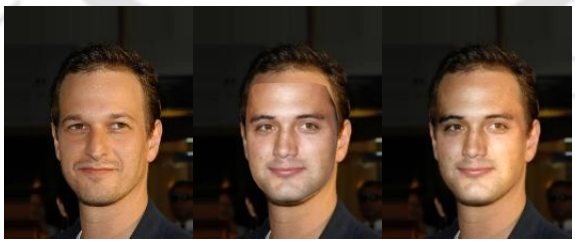


(a) (b)

Figure 3.4: (a) Convex hull of facial landmark points and (b) Triangles using Delaunay Triangulation

(a) (b) (c)

Figure 3.5: (a) A triangle of the source face, (b) A triangle of the destination face and (c) Warped triangle after affine transformation



(a) (b) (c)

Figure 3.6: (a) The destination image, (b) Face swap before blending and (c) Final face swap image

approximates the 3D information of faces. For this, we segment the face of the source image into many triangles using Delaunay Triangulation, as shown in Fig. 3.4b.

2. Step (1) is repeated for the destination face. The advantage of triangulation is that one can transfer each triangle of the source face to a triangle of the destination face while maintaining the proportions between the faces. The matching of two face triangles is done based on the indices of the triangles of the faces. However, directly matching the triangles is not appropriate if the size of the source face is smaller than the destination face or vice-versa.

3. Warped triangles are formed between the points of two triangles using affine transformation to maintain the proportions between the triangles of the source face and the destination face. This is shown in Fig. 3.5.

4. Then, the destination face is replaced with the source face. When the source face is swapped with the destination face, the destination face does not look natural due to variations in color and lighting conditions, as shown in Fig. 3.6b. Formally, the destination face Y is replaced with the source face X using the following steps:

   (a) For each triangle in the destination face Y, the coordinates are computed using Eq. 3.1:

$$\begin{bmatrix} Y_{a,x} & Y_{b,x} & Y_{c,x} \\ Y_{a,y} & Y_{b,y} & Y_{c,y} \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3.1}$$

   Here, the coordinates are given by $[\alpha, \beta, \gamma]^T$. In order to compute this,

the inverse of the 3×3 matrix of each triangle is found using Eq. 3.2:

$$\begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = Y_{\Delta}^{-1} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3.2}$$

Based on the values of α, β and γ, the point is inside triangle if $\alpha \in [0, 1]$, $\beta \in [0, 1]$, $\gamma \in [0, 1]$, and $\alpha + \beta + \gamma \in [0, 1]$.

(b) Using these obtained coordinates, the pixel indices of image X are computed using Eq. 3.3:

$$\begin{bmatrix} x_X \\ y_X \\ z_X \end{bmatrix} = X_{\Delta} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} \tag{3.3}$$

where

$$X_{\Delta} = \begin{bmatrix} X_{a,x} & X_{b,x} & X_{c,x} \\ X_{a,y} & X_{b,y} & X_{c,y} \\ 1 & 1 & 1 \end{bmatrix} \tag{3.4}$$

Then, we compute $[x_X, y_X, z_X]^T$ and convert it into homogeneous coordinates as given in Eq. 3.5:

$$x_X = \frac{x_X}{z_X} \quad and \quad y_X = \frac{y_X}{z_X} \tag{3.5}$$

(c) Now, the pixel value at location (x_X, y_X) is copied back from the image X to the image Y.

5. Finally, seamless cloning is applied to make the face swap more realistic as shown in Fig. 3.6c.

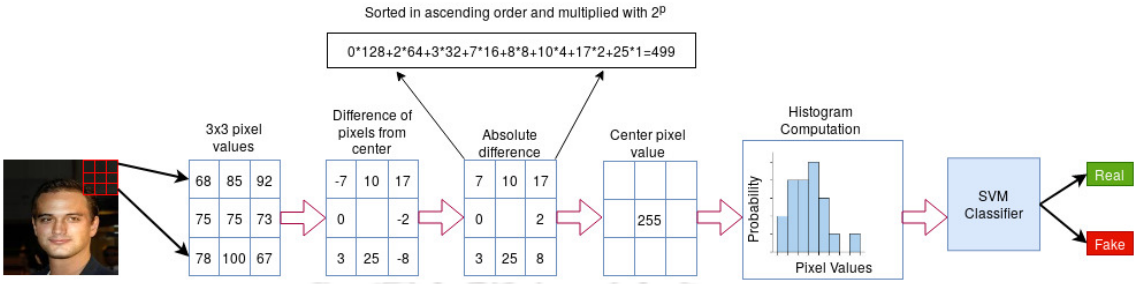


Figure 3.7: The procedure of the proposed FSAD approach

3.3 Proposed Face Swap Attack Detection

In general, digital manipulations perform blending and smoothing to minimize the irregularities because of the differences in the source image. Sometimes more than 90% of the texture surfaces are uniform. In such cases, WLMP feature descriptors that encode the differences between a center pixel and its neighbors are effective in highlighting the most affected regions of images when they are being swapped or switched. It gives more weightage to the closest pixel to the center than the other pixel values. That is, it assigns the weight inversely in proportion to the difference values from the center pixel instead of binarizing them. Moreover, these features retain high-frequency information while reducing low-frequency information. We proposed FSAD for the detection of face swap attacks which is based on WLMP [56] and SVM [103]. Fig. 3.7 shows the steps involved in the proposed FSAD approach. In a nutshell, it is as follows:

1. The input image is segmented into multiple blocks of 3×3 size.

2. For each block, the difference in pixel values from the center and their absolute differences are computed.

3. Since there are eight neighborhood pixels, there are eight pixel difference values. The difference pixel values are sorted in ascending order and multiplied

with 2^p, where $p = 0, ..., 7$ for 8 neighborhood pixel values. The motivation is to give higher weightage to the pixel value which is similar to the center pixel value. Then the obtained final value is mapped to a value between 0 and 255. If the final value is greater than 255, it is set to 255.

4. A histogram feature vector is computed.

5. Finally, the feature vectors extracted from the training dataset are provided to SVM for detecting the presence of the face swap attack.

3.4 Experimental Results

3.4.1 Dataset

A large-scale face attributes dataset, called **CelebFaces Attributes Dataset (CelebA)** [111] is used to evaluate the proposed system. It contains 2,02,599 celebrity face images with 40 attribute annotations each. It covers images of large variations in pose and background clutter. It also contains face images with large diversities and rich annotations.

3.4.2 Performance of the Proposed Face Swapping

The proposed system is trained and tested on the CelebA dataset. Examples of the face swap results on the CelebA dataset are shown in Fig. 3.8. Despite variations in lighting and skin-tone, the proposed system swaps the faces well and obtains results that look realistic.

Examples of the face swap results on the CelebA dataset are shown in Fig. 3.8. Despite variations in lighting and skin-tone, the proposed system swaps the faces well and obtains results that look realistic.

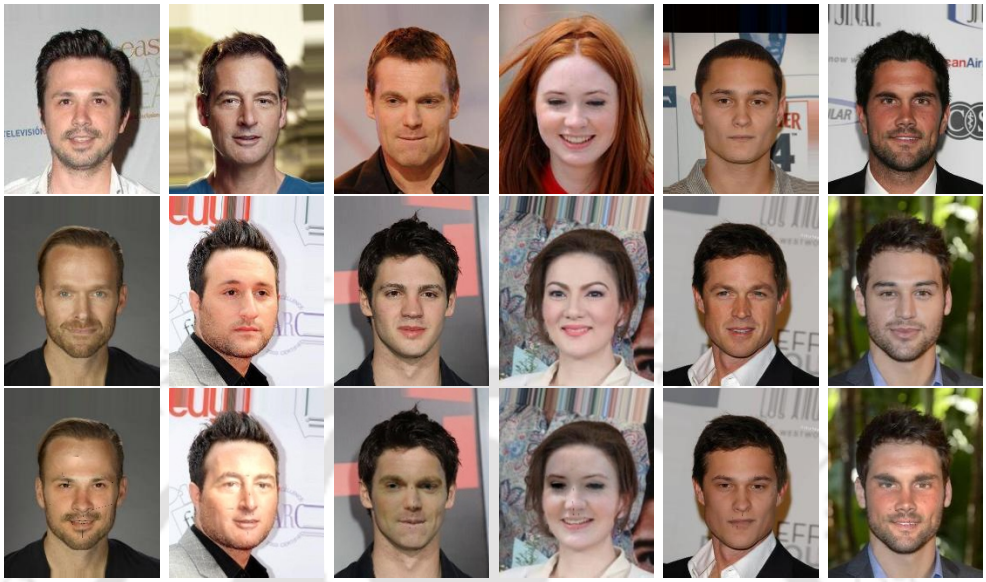


Figure 3.8: Source image, destination image and face swap image (from top to bottom in each column)

The performance of the face swap approach with extracted augmented 81-facial landmarks is compared with face swapping based on 68-facial landmarks. As the augmented 81-facial landmarks cover landmarks on the forehead, face swap results look more realistic and unedited when compared with 68-facial landmarks, as shown in Fig. 3.9. It is also observed that the skin color of the forehead and swapped regions of the face image are not matched when the face swap is done with 68 facial landmarks. Thus, augmented 81 facial landmarks achieve full face swapping, including the forehead and the results appear realistic.

However, the proposed approach for face swapping fails to accurately swap two face images in the following cases: i) if there are large variations in skin-tone, ii) if there are large variations in feature size and iii) if the seam crosses the non-skin regions. The incorrect results for face swapping are presented in Fig. 3.11. We also presented few cases where face swap results based 68-facial landmarks appear more realistic than face swap results based on 81-facial landmarks due to large variations

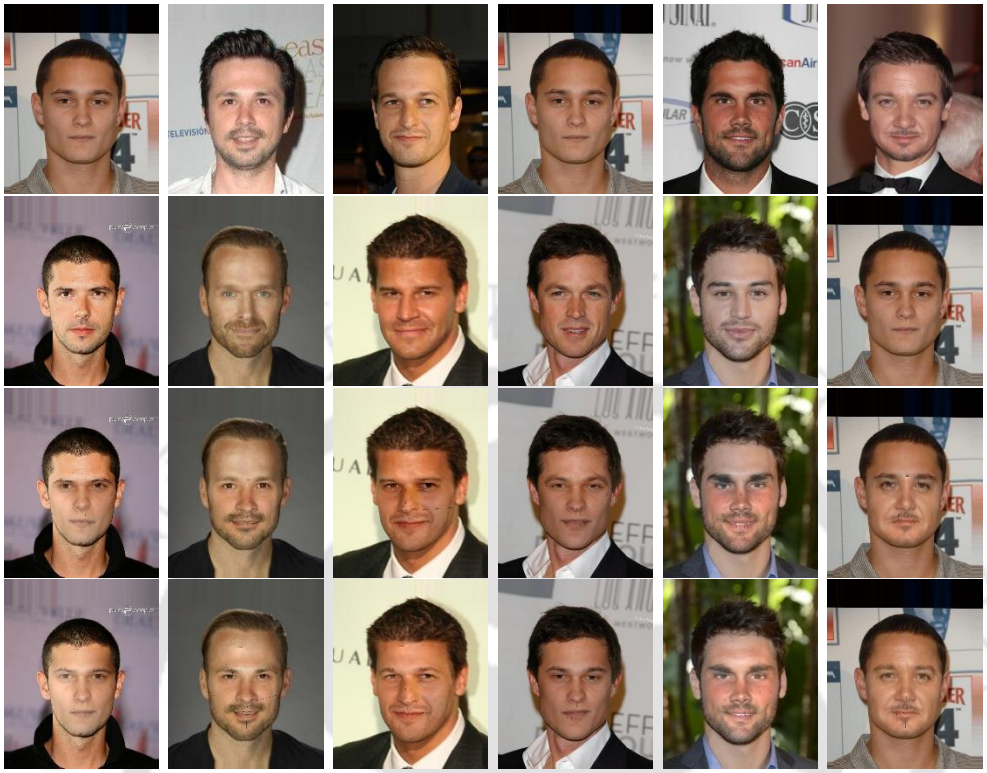


Figure 3.9: Source image, destination image, face swap with 68-facial landmarks and face swap with augmented 81-facial landmarks (from top to bottom in each column)

Figure 3.10: Source image, destination image, face swap with 68-facial landmarks and face swap with augmented 81-facial landmarks in row-wise

in pose, style and lighting conditions as shown in Fig. 3.10.

3.4.3 Performance of the Proposed FSAD

The performance of the FSAD approach is evaluated on different types of SVMs, such as Linear, Polynomial, Sigmoid and Gaussian. Both Sigmoid-SVM and Gaussian-SVM achieve the same values for precision and recall for both fake and original images with an accuracy of 74%, whereas Linear-SVM and Polynomial-SVM achieve a precision of 96% and a recall value of 94% for fake images with a detection accuracy of 95%. Therefore, both Linear-SVM and Polynomial-SVM outperform Sigmoid-SVM and Gaussian-SVM in terms of all parameter values. The detailed FSAD results are shown in Table 3.1.

Figure 3.11: Incorrect face swaps: Source image, destination image and face swap image (from top to bottom in each column)

Table 3.1: Overall face swap attack detection results

S.No	Classifier	Precision	Recall	F1-score	Accuracy (%)
1	Linear-SVM	0.96	0.94	0.95	**95**
2	Polynomial-SVM	0.96	0.94	0.95	**95**
3	Sigmoid-SVM	0.78	0.68	0.73	74
4	Gaussian-SVM	0.78	0.68	0.73	74

3.5 Conclusions

In this chapter, we propose an approach for face swapping based on augmented 81-facial landmark points that also cover the facial landmarks on the forehead. In this method, a procedure for FSAD based on WLMP feature descriptors and SVM is also incorporated. The WLMP feature descriptors are effective in differentiating the high-frequency and low-frequency information when images are attacked. The performance of the proposed system is demonstrated on a real-world dataset. Experimental results show that the proposed system effectively performs swapping of face images despite the images under different variations in pose, lighting and skin-tone. The results look unedited and more realistic. We also present a few cases where the proposed face swap approach fails to swap the faces accurately. For the detection, different types of SVMs such as Linear-SVM, Polynomial-SVM, Sigmoid-SVM and Gaussian-SVM are used to evaluate the performance of the proposed system. Both Linear-SVM and Polynomial-SVM outperform the remaining with a detection accuracy of 95%.

In this chapter, we assumed the attacker swaps the face of one facial image with the face of another facial image to create an attack. In the next chapter, we investigate an attack where the attacker uses the same image to create an attack.

Chapter 4

Detection of Copy-Move Forgery Attacks

In CMF attack, the attacker copies one or more regions of the image and pastes them into one or more parts of the same image and creates a forgery image. It is one of the most actively investigated attacks in digital image forensics. This type of attack is used to either hide or emphasize important objects of the original image and creates one more tampered image intentionally. It is a widely used attack to create tampered images because it is more complicated to detect as some features of the forged region such as noise and color are highly similar to the remaining regions of the image. CMFD refers to the detection of which regions of the image have been forged and to the authentication of digital images.

This chapter proposes a method for detecting CMFs in digital images. It is based on SURF [21] and BRISK [112] features. We first extract feature descriptors using SURF and BRISK. The fused feature descriptors are matched using Hamming distance. Then, we group the matched features into clusters using the HAC technique to reduce false matches, improving the proposed system's accuracy rate and execution time.

The contributions of the chapter are as follows:

- A new detection method for CMF is proposed based on SURF and BRISK features. Euclidean distance is used for matching the fused features to detect forgeries with various geometric transformations.
- We perform a clustering technique to reduce false matches that improve the system's performance.
- Finally, the proposed method is tested on various real-world CMF datasets such as MICC-F220, MICC-F2000, MICC-F8multi [1] and experimental results are presented.

The remainder of the chapter is organized as follows: A brief review of extraction of feature descriptors is presented in Sections 4.1, 4.2 and 4.3, respectively. The proposed method is explained in Section 4.4. The results of the proposed method are elaborated in Section 4.5. We conclude the chapter in Section 4.6.

4.1 SURF Features

This section presents a review on the extraction of SURF features.

4.1.1 Integral Image

Two-dimensional image features can be calculated quickly using integral images. Given an integral image, the sum of pixel values within a region of the image at a point (x, y) can be computed in constant time. Therefore, it improves the performance in terms of computation speed. Its value is calculated from the above and to the left of (x, y). Fig. 4.1 shows an integral image with the rectangular region

[1] www.lambertoballan.net/research/image-forensics/

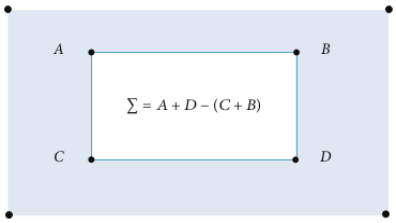


Figure 4.1: Integral image calculation by rectangular region

whose vertices are A, B, C, and D. The sum of the pixel intensities is calculated by the formula, which is written in the rectangular region. Given an input image I and a point (x, y), the integral image $I_{\sum}$ is calculated by the sum of the pixel values above and to the left of the point (x, y). It is formulated by Eq. 4.1:

$$I_{\sum}(x, y) = \sum_{i=0}^{i \leq x} \sum_{j=0}^{j \leq y} I(i, j) \tag{4.1}$$

SURF uses rectangular filters to extract features from images. These features can be rapidly computed using these integral images. With $I_{\sum}$ calculated, the integral image takes only four additions to compute the sum of the pixel values within a rectangular region, independent of its size.

4.1.2 Keypoints Detection

Scale spaces are implemented as image pyramids. In order to get the higher level of the pyramid, a Gaussian is repeatedly applied to smooth the images and subsequently sub-sampled. That is, Laplacian of Gaussian is approximated with a box filter and convolution is used with varying size box filters to create the scale space. Once the scale space is constructed, the Hessian matrix is used to find the extremum point. The determinant of the Hessian matrix is used to decide whether the Eigen values are positive or negative. If the determinant of the Hessian matrix is positive, that means both the Eigen values are either positive or negative. These will be considered as extrema, in case of a positive response. Otherwise, the points will be

ignored. The Hessian matrix $H(X, \sigma)$ in X at scale σ, given a point $X = (x, y)$ in an image I, is defined by Eq. 4.2.

$$H(X, \sigma) = \begin{bmatrix} L_{xx}(X, \sigma) & L_{xy}(X, \sigma) \\ L_{xy}(X, \sigma) & L_{yy}(X, \sigma) \end{bmatrix} \tag{4.2}$$

where $L_{xx}(X, \sigma)$ is the convolution of the Gaussian second order derivative $\frac{\partial^2}{\partial x^2} g(\sigma)$ with the image I in point X and similarly for $L_{xy}(X, \sigma)$ and $L_{yy}(X, \sigma)$. These derivatives are called Laplacian of Gaussian. The approximate determinant of the Hessian matrix is calculated by Eq. 4.3.

$$det(H_{approx}) = D_{xx}D_{yy} - (0.9D_{xy})^2 \tag{4.3}$$

4.1.3 Orientation Assignment

A reproducible orientation for the interesting point is identified to make it rotation invariant. Then, Haar wavelets are computed in x and y directions in the circular neighborhood of a particular radius around keypoint. These increase the robustness and decrease the computational cost. The maximum value is chosen as a dominant orientation for that particular point.

4.1.4 Feature Descriptor Generation

In feature descriptors generation, a square region is firstly constructed around the keypoint, the keypoint is taken as the center point. This square region is again divided into 4×4 smaller sub-regions. Haar wavelet responses are computed for each sub-region. Here, d_x termed as horizontal response and d_y as vertical response. Four responses are considered for each of these sub-regions by Eq. 4.4.

$$V_{subregion} = [\sum d_x, \sum d_y, \sum |d_x|, \sum |d_y|] \tag{4.4}$$

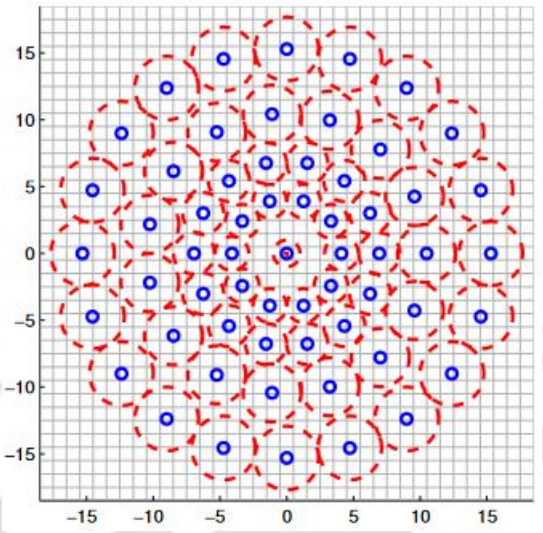


Figure 4.2: The BRISK sampling pattern with $N = 60$

4.2 BRISK Features

BRISK is a binary feature description technique [112]. It is robust against scale transformations. It uses a sampling pattern for feature selection as shown in Fig. 4.2. In Fig. 4.2, the blue dots denote the location and the red-colored dashed lines indicate the radius r_i, which is based on the Gaussian kernel to smooth intensity values of the sampling point for avoiding aliasing at point n_i in the pattern. Thus, the local gradient can be computed by Eq. 4.5:

$$g(n_i, n_j) = (n_j - n_i) \times \frac{(I(n_j, \sigma_j) - I(n_i, \sigma_i))}{\|n_i - n_j\|^2} \tag{4.5}$$

where $g(n_i, n_j)$ is local gradient, $(n_i - n_j)$ denotes the sampling point pairs, $I(n_i, \sigma_i)$ and $I(n_j, \sigma_j)$ represent the smoothed intensity values. BRISK uses sampling pattern around keypoint k rotated by $\alpha = arctan2(g_x, g_y)$ to identify the scale invariance features. The bit vector d_k is calculated by comparing all the short-distance

sampling-point pairs $(n_i^\alpha, n_j^\alpha) \in S_s$ such that every bit b is either 0 or 1 as shown in Eq. 4.6. The size of bit-vector d_k is $64 \times N$ with keypoint in the range specified by $[\delta_{minimum}, \delta_{maximum}]$.

$$b = \begin{cases} 1, & I(n_j^\alpha, \sigma_j) > I(n_i^\alpha, \sigma_i) \\ 0, & \text{otherwise} \end{cases} \tag{4.6}$$

4.3 Keypoints Clustering

We fuse both features extracted using BRISK and SURF. The fuse features are then matched using Hamming distance and grouped into clusters using HAC [113] to reduce false positive matches. The spatial coordinates are considered for calculating the distance between the matched points. It starts clustering the keypoints by considering each keypoint as a cluster initially. Then, it calculates all the reciprocal spatial distances among clusters to find the closest pair of clusters and merges them into a single cluster if they are dissimilar. Such a process is iteratively repeated until there is only one cluster left by the linkage method adopted or the dissimilarity criterion is unsatisfied. Thus, Centroid, Ward and Single linkage methods are used for creating a hierarchy of clusters that can be represented by a tree structure.

4.4 Proposed CMFD Method

The proposed method is based on SURF [21] and BRISK [112] to detect the keypoints and extract their feature descriptors. The matching is performed between the feature descriptor of one keypoint and the feature descriptor of another keypoint to detect CMFs in the image. On matched keypoints, we use clustering to form the clusters and reduce the false positives. Matching the keypoints is done using Euclidean distance and we then detect multiple CMFs in the digital image. Fig.

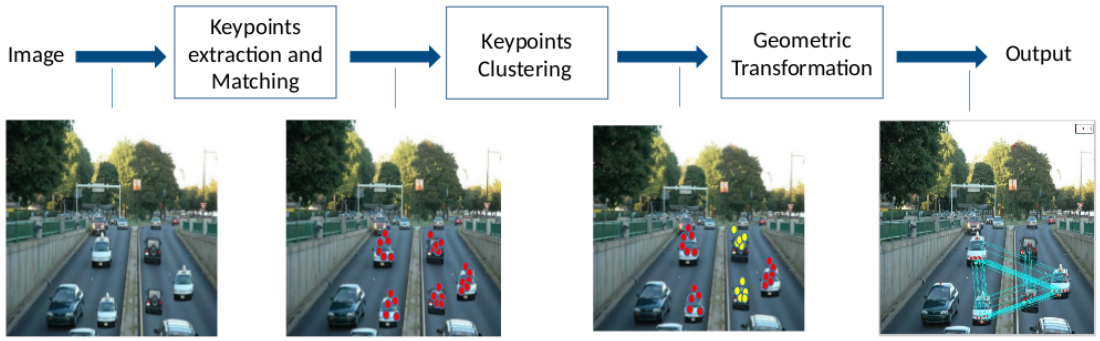


Figure 4.3: An overview of the proposed CMFD system; features extraction, clusters and detection results

4.3 shows a schematic diagram of the proposed CMFD. In the first step, keypoints are detected and keypoints matching is done. Second, clustering is performed on matched keypoints. Finally, detects the forged regions if forgery is presented. An algorithm to detail a step by step process of the proposed method for detecting copy-move forgery attacks is presented in Algorithm 1.

4.5 Experimental Results

4.5.1 Dataset

We evaluate the proposed approach on three different datasets.

First, on MICC-F220 by properly set the threshold t. It is a small dataset and composes of 220 images of resolution varying from 722×480 to 800×600 pixels. Out of which 110 images are originals and the remaining 110 images are forged. On average, the size of the forged area is 1.2% of the whole image.

Second, on MICC-F2000, the evaluation is performed by testing the robustness of the system against different kinds of forgery images. It is a larger dataset and consists of 2000 images 2048×1536 pixels. The forgery area on average is 1.12% of

Algorithm 1 The proposed CMFD based on SURF and BRISK Features

Input: Image

Output: Detected forged regions with image

1. Convert the image into gray-scale image if RGB image.
2. Extract keypoints using SURF from an image $(1, 2, 3, ..., M)$ and for each keypoint, features are extracted $(f_1, f_2, f_3, .., f_M)$.
3. Extract keypoints using BRISK from an image $(1, 2, 3, ..., M)$ and for each keypoint, features are extracted.
4. Fuse both extracted features.
5. For each feature descriptor, matching is performed with each other feature descriptor.
6. If a match exists, clustering is performed using HAC. Euclidean distance is utilised to compute the distance between matched feature descriptors.
7. A line is drawn between the matched objects of different clusters.
8. Forged regions are shown from clusters.

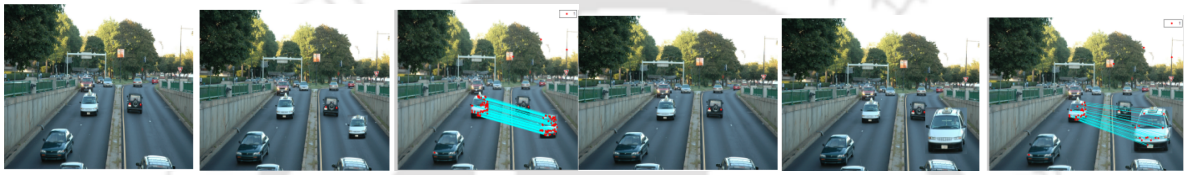
the image.

Third, on a small dataset, MICC-F8multi, contains 8 images in which multiple portions were copied and pasted. In all datasets, the tampered images are obtained by randomly copying region(s) of the image (rectangular or square) and pasting over the image after having applied different attacks such as scaling, rotation and translation.

4.5.2 Evaluation Metrics

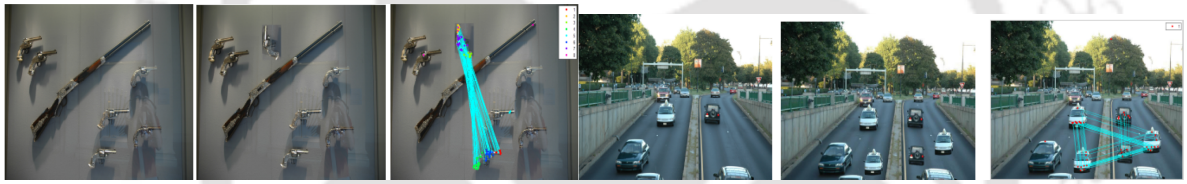
The performance of the system is measured based on TPR, FPR and detection time complexity where

$$TPR = \frac{\text{total images detected as forged being forged}}{\text{total number of forged images}} \tag{4.7}$$



(a) Down scaling attack

(b) Up scaling attack



(c) Rotation attack

(d) Multiple cloning attack

(e) Multiple cloning and scaling attack

Figure 4.4: Performance of the proposed method against various CMF attacks

Table 4.1: Performance of the proposed CMFD in terms of TPR and FPR on varying threshold t

t	Centroid		Single		Ward	
	TPR %	*FPR %*	*TPR %*	*FPR %*	*TPR %*	*FPR %*
0.07	87	8.43	87.3	8.34	87.3	8.34
0.08	92.1	8.5	93.2	8.48	93.2	8.5
0.09	**98**	**7.5**	97.4	7.6	**98**	**7.5**

$$FPR = \frac{\text{total images detected as forged being original}}{\text{total number of original images}} \tag{4.8}$$

We tested the proposed method on all three datasets which cover different CMF attacks on images for checking the robustness of the system. The performance against different types of CMF attacks is shown in Fig. 4.4. Table 4.1 presents the accuracy of the system in terms of TPR and FPR in varying threshold t for the detection of forgeries in digital images.

Table 4.2: Performance comparison of the proposed CMFD

Method	**TPR**	**FPR**	**Time(s)**
Fridrich et al. [114]	89	84	294.69
Popescu and Farid [62]	87	86	70.97
Amerini et al. [2]	100	8	4.94
Mishra et al. [77]	73.64	3.64	2.85
Our method	**98**	**7.5**	**6.5**

Table 4.2 shows the performance of the system and comparison with other methods. The proposed method achieves 98% of TPR and outperforms SOTA methods in the presence of multiple cloning.

4.6 Conclusions

We propose a detection method for CMF to support image forensics investigation based on SURF and BRISK features. The proposed method was evaluated on three benchmark datasets. Given a forged image, it can efficiently detect if certain portions have been tampered with. The results show its effectiveness against CMF attacks with respect to various geometric transformations.

In the next chapter, we investigate a digital image attack where an attacker adds small perturbations to an input image to create attacks.

Chapter 5

Defense Methods against Facial Adversarial Attacks

DL models have been gaining popularity in various ML tasks because they can represent higher-level concepts from low-level features and generalize them involving a highly complex input space such as image classification, etc. However, these models are vulnerable to versions of input images with slight perturbations added intentionally to the input image to cause misclassification [10]. Such attacks are known as adversarial attacks. These attacks are generated intentionally to mislead the classifier for a wrong prediction. Based on the assumption of the attacker's knowledge, there are various types of adversarial attacks where each attack with a specific goal. In general, the primary goal is to disrupt the input image by adding small perturbations to cause the desired misclassification by the classifier.

Based on assumptions of the attacker's knowledge, attacks are broadly classified into white-box and black-box attacks. In a white-box attack, it is assumed that the attacker is fully aware of the model's inputs, outputs, weights and architecture. In a black-box attack, it presumes the attacker is just aware of the model's inputs and outputs and is unaware of its underlying architecture or weights. While generating

adversarial attacks, attackers can have a variety of goals such as source/target misclassification and misclassification. The goal of misclassification is the attacker does not care what the new classification is but only wants the output classification to be wrong. A source/target misclassification means the adversary tries to change an image from a certain source class so that it is classified as a particular target class.

In this chapter, we investigate two effective defense methods for detecting facial adversarial attacks. The first method is against intensity-based facial adversarial attacks and another method is against geometry-based adversarial attacks. Distinctive feature analysis is utilized for each method and linear and non-linear classifiers to effectively detect facial adversarial attacks from the clean images. The following is the list of key contributions of this chapter:

- Two defense methods are proposed against facial adversarial images. Intensity-based and geometry-based facial adversarial images are generated based on P-FGSM and FLM, respectively.
- WLMP features are extracted for detecting intensity-based adversarial noises in face images.
- ELA is performed for the detection of geometry-based adversarial noises in face images.
- Experimental results are presented.

The remainder of the chapter is organized as follows: Section 5.1 presents the problem formulation. The proposed defense method against intensity-based facial adversarial attacks along with a brief review on related works is discussed in Section 5.2. Its experimental results are elaborated in Section 5.2.3. The proposed defense method against geometry-based facial adversarial attacks along with related works

is discussed in Section 5.3. Its performance is detailed in Section 5.3.4. We conclude the chapter in Section 5.5.

5.1 Problem Definition

We study the defense method against non-targeted adversarial facial attacks. Let $X' = [0,1]^{R \times C \times Ch}$ be the input image space, where R is the number of rows, C is the number of columns and Ch is the number of channels. Given an image classifier $C(.)$ and a source image $x \in X'$, a non-targeted adversarial image of x is a small perturbed image $x^{adv} \in X'$ such that $C(x) \neq C(x^{adv})$ and $D(x, x^{adv}) \leq t$ for some dissimilarity function $D(.,.)$ and $t \geq 0$. Given a set of N clean images $x_1, ..., x_N$ and a target classifier $C(.)$, an adversarial attack aims to generate $x_1^{adv}, ..., x_N^{adv}$ where each x_n^{adv} is an adversarial image for x_n. On the other way, a defense method aims to make the prediction on adversarial image $C(x^{adv})$ that is similar to the prediction on the corresponding clean image.

5.2 Intensity-based Facial Adversarial Attacks

Intensity-based adversarial attack directly tries to alter the intensity of an input image. This section describes the proposed defense method against intensity-based facial adversarial attacks. First, face adversarial attacks are generated based on P-FGSM [115]. It picks the target class label from a subset of adaptive class labels. Thus, it reduces the probability of deducing the mapping between the target and original classes. As a result, the adversarial image is more protected and increases the probability of misleading the classification by a classifier. Next, the encoded WLMP features are extracted from an input image and provided to various types of SVM classifiers to discriminate between adversarial examples and real ones. The effectiveness of the proposed defense method is shown on a real-world face image

dataset.

5.2.1 Adversarial Attacks Generation

A simple and effective method, FGSM, for creating intensity-based adversarial attacks is presented in [10]. FGSM is a type of white-box attack with the goal of misclassification. It uses gradients in the way the NNs learn the gradients to attack them. It adjusts the input data to maximize the loss on the same backpropagated gradients instead of minimizing the loss by adjusting the weights using the backpropagated gradients. This can be summarized using the following eq. 5.1:

$$Adv\_X = X + \epsilon \times sign(\nabla_X J(\theta, X, Y)) \tag{5.1}$$

where

Adv_X: Adversarial image

X: The input original image

Y: The true label of the input image

ϵ: Small value to multiply the signed gradients to control perturbations such that they are small enough to deceive the human eye but large enough to mislead the neural network

θ: Neural network model parameters

J: The loss function

An example for FGSM attack is as shown in Fig. 5.1. Where X depicts an input image and it is rightly classified as "macaw" with 97.3% confidence score. A small perturbation $\epsilon \times sign(\nabla_X J(\theta, X, Y))$ is added to the input image. The obtained adversarial image $X + \epsilon \times sign(\nabla_X J(\theta, X, Y))$ is wrongly classified as "bookcase" with 88.9% confidence score, when it still appears as "macaw". Iterative FGSM (I-FGSM) [116] is an extended version of FGSM. It iteratively generates perturbations until a maximum number of iterations or desired misclassification probability is

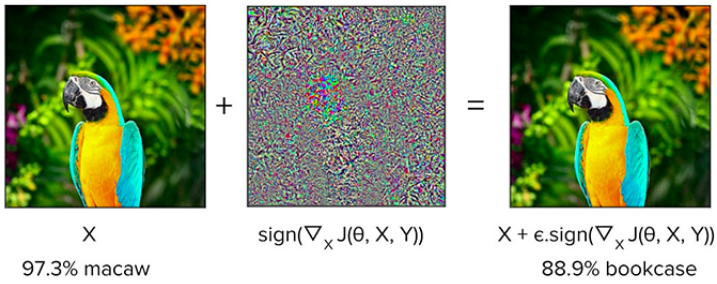


Figure 5.1: The FGSM for adversarial image generation [10]

reached. The final $X' = X_N$ is obtained as:

$$X' = X'_{N-1} + \epsilon \times sign(\nabla_X J(\theta, X'_{N-1}, Y)) \tag{5.2}$$

from the initialization $X'_0 = X$.

We generate intensity-based adversarial images based on P-FGSM [115]. It is an iterative FGSM with the goal of target misclassification. Let an image I and $\hat{y}_i$ be its true class label of one of the scene types shown in I. Let a set of N scene classes of an image be $y_1, ..., y_i, ..., y_N$. Then, a multiclass classifier M is applied to image I to generate a one-hot vector y of size N-dimensional, which is given by:

$$y = M(I) \tag{5.3}$$

where $y = \{y_1, ..., y_i, ..., y_N\}$ is obtained from a selection on the probability vector $p = \{p_1, ..., p_i, ..., p_N\}$. Here, p_i is the probability of the scene class y_i of the image I.

$$p_i = p(y_i/I) \tag{5.4}$$

A transformation T is defined such that $\hat{I} = T(I)$ to induce M to classify the image I with a different scene label:

$$y \neq M(\hat{I}) \tag{5.5}$$

The transformation T applies distortion to the image I. The distortion should be minimal so that T is unnoticeable. Moreover, T should not be reversible such that the true class $\hat{y}_i$ can not be inferred from the predicted class $M(\hat{I})$ or from the probability distribution of the predicted classes. Thus, T is defined as follows:

$$\hat{I} = T(I) = I + \delta_I^* \tag{5.6}$$

where δ_I^* is an adversarial perturbation. It is generated as follows:

$$\delta_I^* = arg_{\delta_I} max J_M(\theta, I + \delta_I, y) \tag{5.7}$$

P-FGSM generates adversarial images by adaptively targeting a class label $\hat{y}$ that is picked as a function of the classification probability vector p. It achieves a high misclassification rate by utilizing the fact that the true class labels are always among the class labels with the highest collective probabilities. Let $p' = \{p'_1, ..., p'_N\}$ are the elements of p that are sorted in non-increasing order. P-FGSM arbitrarily picks $\hat{y}$ from the subset of classes if its cumulative probability crosses a specified threshold $\sigma \in [0, 1]$:

$$\hat{y} = R(\{y_j : \sum_{i=1}^{j-1} p'_j > \sigma\}), \tag{5.8}$$

where R is a function that selects a class label arbitrarily from the input set and σ is a threshold to control the number of classes to select $\hat{y}$: a higher σ denotes a smaller subset of target classes. P-FGSM generates the adversarial image $\hat{I} = \hat{I}_N$ iteratively, starting from $\hat{I}_0 = I$, as

$$\hat{I} = \hat{I}_{N-1} - \epsilon \times sign(\Delta_I J_M(\theta, \hat{I}_{N-1}, \hat{y})), \tag{5.9}$$

by increasing the prediction probability of class label $\hat{y}$ until the desired classification probability or a threshold on the maximum number of iterations is reached.

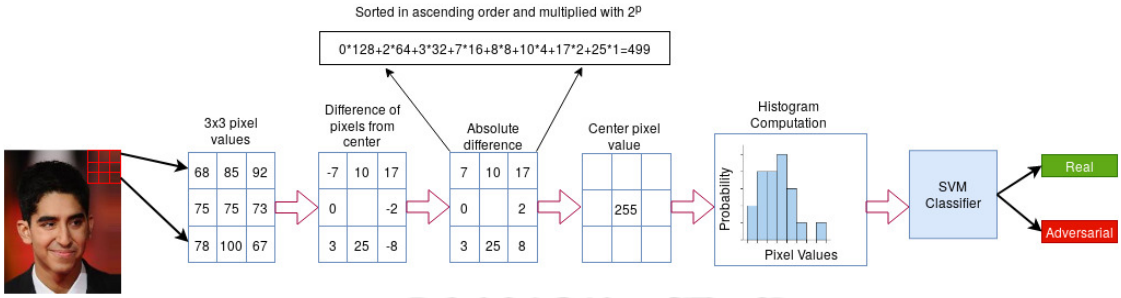


Figure 5.2: The procedure for proposed defense method

5.2.2 Defense Method

Although there have been several methods for defending intensity-based facial adversarial attacks, it is a challenging task and remains an active problem. Since small perturbations added to the input image by many adversarial methods appear like high-frequency noise, several authors have recommended leveraging the benefit of image preprocessing and denoising techniques as a potential defense against adversarial images. There is a lot of variance in preprocessing methods such as applying median filtering and lowering the precision of input data [117] or using JPEG compression [118]. However, such defense methods may be effective against certain attacks, it has been shown that they fail in the white-box case in which the attacker is aware of the defense.

The proposed defense method is based on WLMP features [56] and SVM [103]. The flowchart of the proposed approach is shown in Fig. 5.2. It is explained in detail in section 3.3.

5.2.3 Experimental Results

The CelebA dataset is used to train and test the proposed method. It generates adversarial images based on P-FGSM. Examples of the experimental results obtained using the proposed system on a real-world dataset CelebA are shown in Fig. 5.3.

Figure 5.3: Original images (first row) and adversarial images generated with P-FGSM (second row)

The performance of the proposed defense method is demonstrated by training and testing on different types of SVMs. It effectively defended intensity-based adversarial images from the original images with an accuracy of 98.75% under Linear SVM. The experimental results are shown in Table 5.1.

Table 5.1: Performance of the proposed defense method

Classifier	**Precision**	**Recall**	**F1-score**	**Accuracy(%)**
Linear-SVM	1.0	0.975	0.987	**98.75**
Poly-SVM	1.0	0.97	0.984	98.5
Gaussian-SVM	1.0	0.96	0.979	98
Sigmoid-SVM	0.72	0.725	0.723	72.2

5.3 Geometry-based Facial Adversarial Attacks

The geometry of the face is distinctive for each person. It offers highly discriminating inputs for face recognition. In this section, we discuss the proposed defense method against geometry-based facial adversarial attacks. First, we generate adversarial

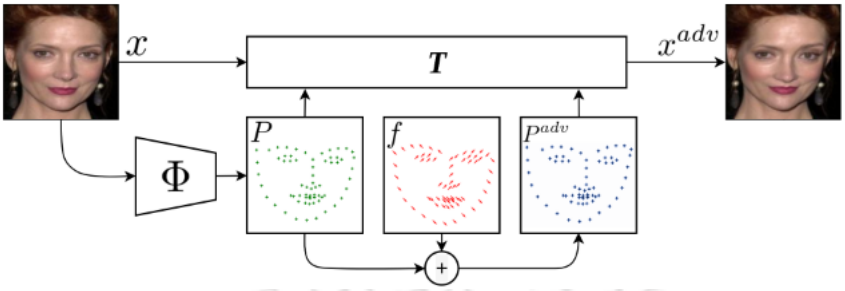


Figure 5.4: Geometry-based adversarial face image generation by FLM [11]

images based on the FLM method. Next, ELA is performed and features are extracted from the input facial images. Then, the extracted features are provided to different types of classifiers for the detection of the adversarial images.

5.3.1 Adversarial Attacks Generation

We generate geometry-based adversarial face attacks based on FLM [11]. It uses the k spatial locations of facial landmarks to create such attacks as shown in Fig. 5.4. It is briefly discussed as follows: Let ϕ be a function for landmarks detection that converts the input image space into a collection of k locations of 2D face landmarks $L = \{l_1, l_2, ...l_k\}$, where $l_i = (p_i, q_i)$. Let $l_i^{adv} = (p_i^{adv}, q_i^{adv})$ is the obtained after the transformation of landmark l_i and it defines the i^{th} landmark location of the corresponding adversarial image x^{adv}. FLM defines flow or displacement field f per landmark to identify the corresponding landmark location of the adversarial image and manipulates the input image space based on L. It optimizes the spatial flow vector $f_i = (\Delta p_i, \Delta q_i)$ for the i^{th} landmark $l_i^{adv} = (p_i^{adv}, q_i^{adv})$. It uses the gradient direction of the prediction like as FGSM [10] to iteratively determine the displacement field f for each landmark. The adversarial landmark l_i^{adv} can be computed from the original landmark l_i and its corresponding displacement vector

f_i as:

$$l_i^{adv} = l_i + f_i, \qquad (p_i^{adv}, q_i^{adv}) = (p_i + \Delta p_i, q_i + \Delta q_i) \tag{5.10}$$

The displacement field f for k 2D landmarks where k is particularly small compared to the pixel count of the input image. The adversarial face image is generated by transforming the original image using the transformation function T, defined as follows:

$$x^{adv} = T(L, L^{adv}, x) \tag{5.11}$$

where T represents a transformation function that maps the landmarks of the input image space L to the adversarial image space L^{adv}.

5.3.2 Error Level Analysis

ELA [119] is a forensic methodology that makes use of the lossy compression methods of forged images to reveal the fake. It works on image grids which are re-compressed independently by a lossy technique with a known error rate. It then calculates the absolute difference between the image suspected of being under attack and the re-compressed image. Formally, ELA is defined as follows: For each color channel error levels, $ELA(r, c)$ where r and c are the indices of row and column respectively, can be denoted by

$$ELA(r, c) = |I(r, c) - I_{recompressed}(r, c)|, \tag{5.12}$$

where I denotes the image suspected of being under attack and $I_{recompressed}$ is the re-compressed image. Total averaged error levels across all color channels is denoted as,

$$ELA(r, c) = \frac{1}{3}\sum_{i=1}^{3} |I(r, c) - I_{recompressed}(r, c)|, \tag{5.13}$$

where i varies from 1 to 3 for an RGB image.

The difference obtained between the two images represents the error levels associated with the pixels of the original and re-compressed images. The error

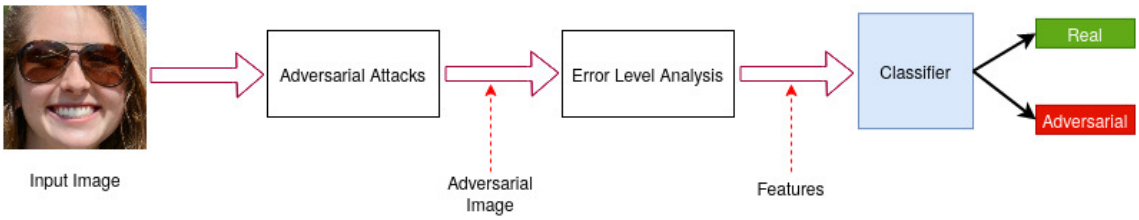


Figure 5.5: The overall procedure of the proposed defense method against geometry-based adversarial attacks

levels shows a magnitude of change that occurred in the image suspected of being under attack. If the magnitude of error levels is small, the pixel has reached its local minima for error at the stated error rate. The pixels are likely to be attacked and not at their local minima if the magnitude of error levels is large.

5.3.3 Defense Method

Once ELA is performed, the extracted features are provided to different types of classification models to detect the adversaries from the clean images. The overall flowchart of our defense method is as shown in Fig. 5.5. The proposed defense is evaluated on various types of classification models such as Logistic Regression (LR), Random Forest (RF), Ada Boosting (AdaBoost), Gradient Boosting (GB), RF with Gradient Boosting (RF-GB), RF with Ada Boosting (RF-Ada) and XGBoost (xgBoost). The experimental results show that the proposed defense effectively classifies adversarial attacks from the original facial images.

5.3.4 Experimental Results

Our experiments evaluate the robustness of the proposed defense on CelebA Dataset [111]. Examples of FLM attacks are shown in Fig.5.6. The performance is evaluated on different types of classifiers as shown in Table 5.2.

If the precision is considered as the efficiency metric of the classifier, the

Figure 5.6: Some of the results for geometry-based face attacks generated using FLM. The first row represents original images from the CelebA dataset

Table 5.2: Overall results of the proposed defense method

S.No	Classifier	Precision	Recall	F1-score	Accuracy (%)
1	Logistic Regression	0.99	1.00	1.00	**99.75**
2	Gradient Boost	0.99	1.00	0.99	99.38
3	Random Forest with Gradient Boosting	0.99	1.00	0.99	99.38
4	Random Forest with Ada Boosting	0.99	1.00	0.99	99.38
5	XGBoost	0.99	0.99	0.99	99.13
6	Ada Boosting	0.97	0.96	0.97	96.77
7	Random Forest	0.97	0.96	0.97	96.77

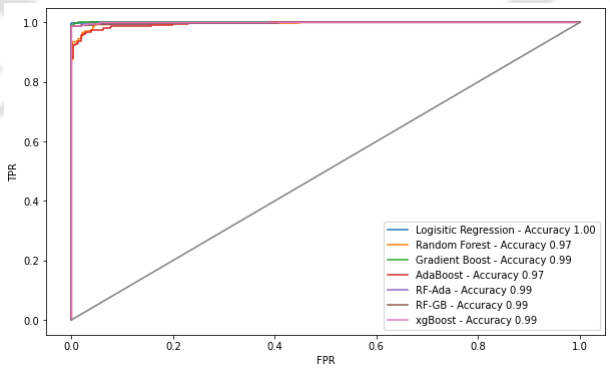


Figure 5.7: Performance comparison of the proposed defense on all classification models

classifiers LR, GB, RF-GB, RF-Ada and xgBoost with precision 0.99 outperform the remaining classifiers. The classifiers RF and AdaBoost have the lowest precision 0.97 while discriminating the adversarial images from the clean images. If the recall is considered as the efficiency metric of the classifier, the classifiers LR, GB, RF-GB and RF-Ada have better recall with 1.00 against the remaining classifiers. In the remaining classifiers, RF and AdaBoost have the lowest recall 0.96. Considering F1-score as the efficiency metric of the classifier, the classifier LR with F1-score of 1.00 outperforms other classifiers. Among the other classifiers, RF and AdaBoost observe the lowest F1-score of 0.97.

The performance comparison of our defense method on all classification models with respect to TPR and FPR is shown in Figure 5.7. Among all the classifiers, LR outperforms the remaining classification models in terms of all metrics with 0.99 precision, 1.00 recall, 1.00 F1-score and 99.75% accuracy. Whereas, RF and AdaBoost have the same and lowest values in all the metrics to detect adversarial face images with 0.97 precision, 0.96 recall, 0.97 F1-score and accuracy of 96.77%.

5.4 Evaluating Robustness of Intensity-based and Geometric-based Adversarial Attacks

Almost all intensity-based attacks augment the input samples with high-frequency components and employ a $l_p - norm$ constraint to regulate the distortion. The adversarial samples may not necessarily sit on the same manifold as the natural samples since the $l_p - norm$ is not a perfect similarity metric. On the other hand, geometric-based adversarial attacks are extremely robust against adversarial training compared to intensity-based adversarial attacks because they are targeting the most important locations in the images using geometric perturbations. We use P-FGSM [115] and FLM [11] for intensity-based and geometric-based adversarial attacks. We

evaluate the robustness of intensity-based and geometric-based adversarial attacks by extracting the encoded WLMP features with various classifiers on CelebA dataset. Geometric-based adversarial attacks are much more resistant against all evaluating classifiers except Sigmoid SVM. The overall statistics for evaluating the robustness of both adversarial attacks are presented in Table 5.3.

Table 5.3: Robustness comparison of intensity-based and geometric-based adversarial attacks on CelebA dataset

S.No	Classifier	Intensity-based Adversarial Attack (P-FGSM)				Geometric-based Adversarial Attack (FLM)			
		Precision	Recall	F1-score	Accuracy(%)	Precision	Recall	F1-score	Accuracy(%)
1	Linear SVM	0.96	0.97	0.96	**96.4**	0.77	0.75	0.76	**76.16**
2	Polynomial SVM	0.89	0.96	0.92	**91.89**	0.75	0.79	0.77	**76.16**
3	Random Forest	0.91	0.98	0.94	**94.14**	0.74	0.8	0.77	**75.99**
4	Sigmoid SVM	0.52	0.54	0.53	**51.35**	0.58	0.63	0.6	**58.77**
5	Gaussian SVM	0.91	0.94	0.92	**92.34**	0.7	0.79	0.74	**72.35**
6	k-NN	0.87	0.99	0.93	**91.89**	0.74	0.49	0.59	**65.73**

5.5 Conclusions

In this chapter, we propose two defense methods against different types of facial adversarial attacks. One method is against a well-protected version of intensity-based facial adversarial attacks. We proposed the encoded feature descriptors and provided to different types of SVMs to effectively differentiate such attacks from the original. The second method is against geometry-based facial adversarial attacks in which adversarial attacks are generated based on the facial landmarks of the facial image. ELA is performed and the performance is evaluated on various types of classifiers to detect such types of attacks. The performance of the proposed defense methods is shown on real-world datasets and the results show their effectiveness in detecting facial adversarial images from the original.

5.5 Conclusions

In the next chapter, we investigate various types of image restoration techniques for improving the facial adversarial robustness of different types of classifiers on other types of adversarial attacks.

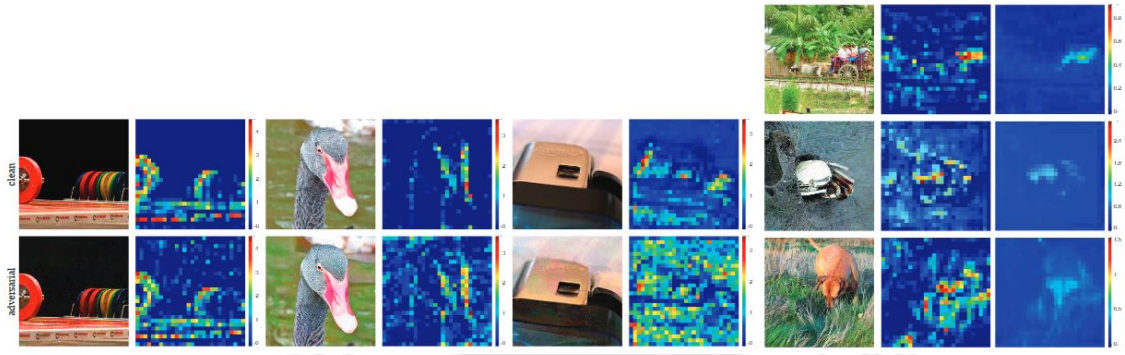
Chapter 6

Image Restoration for Improving Facial Adversarial Robustness

The sensitivity of DL models to adversarial attacks can be problematic and even prevent them from being used in safety and security-critical applications. When human safety is at stake, such as in perceptual tasks for autonomous driving, the problem becomes even more serious. Despite multiple defense methods that have been developed to avoid misclassification by the classifier [82, 95, 120, 121], many of these defenses can be easily attacked by more powerful adversarial examples [30, 122]. Thus, improving adversarial robustness is essential for mitigating the effect of adversarial attacks on classification models.

6.1 Motivation

It is observed that despite adversarial noises being modest in the pixel space, they produce a significant amount of noise in their corresponding feature maps, as shown in Fig. 6.1. That is, the features for the clean image appear to concentrate mostly on semantically informative regions in the image, while the feature maps for the



(a) (b)

Figure 6.1: (a) Feature maps of clean and adversarial images and (b) Feature maps of adversarial images before and after denoising [12]

adversarial images are active across semantically irrelevant regions as well. It is also observed that image de-noising techniques can significantly minimize the amount of noise in the feature maps.

In this chapter, we present an extension of our work [123] by introducing deep image restoration networks for improving adversarial robustness. First, facial adversarial attacks are crafted based on StyleGAN [35], FLM [11] and P-FGSM [115]. The crafted adversarial images are enhanced using deep image restoration networks to bring them back into the original space. Then, we extract the encoded WLMP features from an input image and provide them to various types of SVM, Random Forest (RF) and k-NN classifiers for evaluating the performance. The experimental results show that the proposed model effectively discriminates adversarial images from real ones and significantly improves performance on all the evaluating classification models.

Key contributions of this chapter are listed as follows:

- A new defense method is proposed against facial adversarial images based on deep image restoration networks. It enhances adversarial images by

bringing them back into the original space using Bilateral (BL) filter and Super Resolution (SR) before detecting them from the original. Adversarial attacks are generated using FLM, StyleGAN and P-FGSM methods.

- The encoded features are extracted from each facial image and provided to k-NN, Random Forest and different types of SVM classifiers to detect adversarial facial images from the original.

- Finally, the performance of the proposed defense is demonstrated on real-world datasets and experimental results before and after using image restoration networks are presented.

The remainder of the chapter is organized as follows: The related topics are briefed in Sections 6.2 and 6.3. Section 6.4 presents the proposed model for improving adversarial robustness. Its experimental results are elaborated in Section 6.5. We conclude the chapter in Section 6.6.

6.2 Adversarial Attacks Generation

The adversarial face images are generated based on attacks P-FGSM [115] and FLM [11], as explained in the sections 5.2.1 and 5.3.1 of chapter 5.

6.3 Feature Denoising and Deep Image Restoration Networks

An effective denoising technique can help to mitigate the effect of added perturbations if not eliminated because all adversarial attacks add noise to an input image in the form of well-crafted small perturbations. Image denoising, either in the spatial

or frequency domain, causes a loss of textural information, which is counterproductive to our goal of producing clean image-like performance on denoised images. We denoise adversarial images using the BL filter [124], which is the most used edge-preserving denoising technique. It combines both range and domain filtering to smooth images and preserves edges in a way similar to human performance. It averages only perceptually similar colors and preserves only perceptually visible edges.

SR reconstructs a high-resolution image I^{SR} from a low-resolution image I^{LR}. Depending on the situation, the relationship between I^{LR} and the original high-resolution image I^{HR} can change. Recently, DNNs [125, 126] have been shown to significantly enhance the Peak Signal-to-Noise Ratio (PSNR) in the SR problem. We reconstruct high-resolution images for denoised images based on the Enhanced Deep Super-Resolution (EDSR) network [127]. It consists of residual blocks and ResNet architecture and produced significantly improved performance in the single image SR problem.

6.4 Proposed Model for Improving Facial Adversarial Robustness

Once the adversarial images are restored to the original space, we extract WLMP features [56] from each facial image. Generally, it is observed that smoothing and blending are common digital image editing methods used to remove abnormalities in fake or altered face images. As a result, the texture surfaces sometimes appear more than 90% unaltered. In such cases, the encoded WLMP features are more effective in highlighting the most altered regions of the face images when they are being altered or attacked. The basic idea of the WLMP features is that it encodes the differences computed by a center pixel with its adjacent pixels and weighs the nearest pixel to

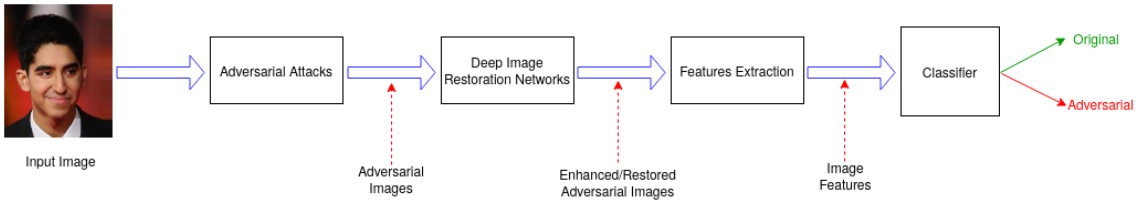


Figure 6.2: The overall procedure of the proposed defense method

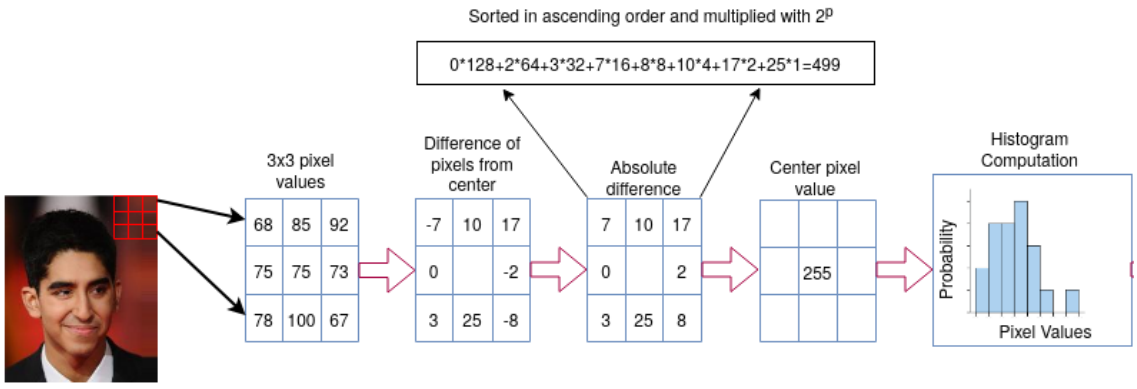


Figure 6.3: The procedure for the extraction of WLMP features

the center pixel more than the far away pixels. It is observed that these features also retain the information of high-frequency pixels while reducing the information of low-frequency pixels. The flowchart shown in Fig. 6.3 describes the procedure for extracting the WLMP features for each facial image. These extracted features are provided to various types of SVM [103] classifiers, RF and k-NN classifiers to discriminate the face adversarial images from the original. The procedure of the proposed model for improving adversarial robustness is as shown in Fig. 6.2.

The steps for extracting WLMP features for each facial image are detailed as follows:

1. The input face image is first divided into multiple blocks of size 3×3.

2. The differences between the center pixel and its adjacent pixels are computed for each block and their absolute values are also computed.

3. Since there are eight adjacent pixels to each center pixel, eight differences are obtained. To give a higher weightage to the pixels that are close to the center pixel, the obtained differences are sorted in increasing order and each absolute value of the difference is multiplied by a value 2^p, where $p = 0, 1..., 7$. Then, the final value of the center pixel is set to a value in the range of 0 and 255. That is, its value is set to 255 if the obtained value is greater than 255.

4. Finally, the histogram feature vector is computed.

Algorithm 2 A Robust Defense against Adversarial Facial Images with Image Restoration (BL + SR)

/* Image de-noising Input */

Input: Adversarial image x^{adv} // P-FGSM, FLM, StyleGAN Attacks

Output: Denoised Image $x_D = D(x^{adv})$

1. Convert the RGB image into gray color image using the transformation $0.299 * R + 0.587 * G + 0.114 * B$.

2. Denoise noisy patterns in the image using BL Filter.

3. Revert the denoised image back to RGB.

/* Image Super-Resolution (SR) */

Input: Denoised image $x_D = D(x^{adv})$

Output: Super Resolved Image $x_{SR} = N(x_D)$

4. Transform adversarial images back to normal image space using deep image restoration networks: N(.).

/* Adversarial Images Detection */

5. Extract WLMP encoded features for the recovered or super resolved images.

6. Forward the extracted features to the classifier model for correct prediction.

Figure 6.4: Examples of original images, adversarial images generated using FLM, restored adversarial images by BL filter and restored adversarial images by BL+SR (from top to bottom in each column)

6.4.1 Algorithm Description

The algorithm of the proposed defense method is provided in Algorithm 2. We used three techniques P-FGSM, FLM and StyleGAN for generating adversarial images. Then, denoising is performed on the adversarial face image using the BL filter. It smoothes the effects of adversarial noise. After that, SR is performed as a mapping function to enhance the visual quality of images, which brings the images in the adversarial space into the original space in high-resolution. Then, encoded WLMP features are extracted for each facial image and trained with different types of SVM classifiers, RF and k-NN. Our defense method minimizes the effect of adversarial perturbations in the image domain and significantly improves the overall performance of the classifier.

Figure 6.5: Examples of original images, adversarial images generated using P-FGSM, restored adversarial images by BL filter and restored adversarial images by BL+SR (from top to bottom in each column)

6.5 Experimental Results

The proposed defense method is trained and tested on two real-world image datasets CelebA [111] and Flickr-Faces-High Quality (FFHQ) [35]. Its performance is demonstrated with different types of classifiers.

6.5.1 FFHQ Dataset

It contains 70,000 real faces from Flickr and 70,000 fake faces generated by using StyleGAN [35]. It contains considerable variation in terms of age, ethnicity and image background. It also has good coverage of accessories such as eyeglasses, sunglasses, hats, etc. Only images under permissive licenses were collected. Various automatic filters were used to prune the set and finally, Amazon Mechanical Turk was used to remove the occasional statues, paintings, or photos of photos.

Table 6.1: Overall results of the proposed defense method on CelebA Dataset with P-FGSM attacks

Image Restoration	Classifier	Precision	Recall	F1-score	Accuracy(%)
No	Linear SVM	1.0	0.98	0.99	98.75
	Polynomial SVM	1.0	0.97	0.98	98.5
	Random Forest	1.0	0.97	0.97	98.5
	Sigmoid SVM	0.72	0.73	0.72	72
	Gaussian SVM	1.0	0.96	0.98	98
	k-NN	0.98	0.97	0.98	97.5
BL+SR	Linear SVM	0.99	1.0	1.0	99
	Polynomial SVM	0.98	1.0	0.99	98
	Random Forest	0.99	1.0	0.99	99
	Sigmoid SVM	0.93	1.0	0.96	93
	Gaussian SVM	0.99	1.0	0.99	99
	k-NN	0.98	1.0	0.99	98

Fig. 6.4 depicts the original facial images, their corresponding FLM attacks, restored adversarial images after BL and restored adversarial images after BL+SR row-wise, respectively. Similarly, Fig. 6.5 shows examples of original face images, their P-FGSM attacks, restored adversarial images after BL and restored adversarial images after BL+SR row-wise, respectively.

Table 6.2: Overall results of the proposed defense method on FFHQ Dataset

Adversarial Attack	Classifier	No Image Restoration				BL				BL+SR			
		Precision	Recall	F1-score	Accuracy (%)	Precision	Recall	F1-score	Accuracy (%)	Precision	Recall	F1-score	Accuracy (%)
StyleGAN	Linear SVM	0.71	0.7	0.7	**70.1**	0.85	0.85	0.85	**80.07**	0.85	0.85	0.85	**80.07**
	Polynomial SVM	0.78	0.78	0.78	**77.94**	0.89	0.89	0.89	**85.29**	0.89	0.89	0.89	**85.29**
	Sigmoid SVM	0.66	0.65	0.65	**65.69**	0.83	0.82	0.82	**77.12**	0.83	0.81	0.82	**76.47**
	Gaussian SVM	0.62	0.56	0.59	**60.29**	0.8	0.71	0.75	**68.63**	0.8	0.71	0.75	**68.95**
	Random Forest	0.66	0.67	0.66	**66.18**	0.83	0.82	0.82	**76.47**	0.83	0.81	0.82	**75.82**
	k-NN	0.71	0.48	0.57	**63.73**	0.88	0.73	0.8	**75.16**	0.88	0.71	0.79	**74.18**
FLM	Linear SVM	0.67	0.58	0.62	**64.66**	0.78	0.66	0.72	**68.09**	0.78	0.67	0.72	**68.45**
	Polynomial SVM	0.65	0.43	0.52	**60.21**	0.77	0.53	0.63	**62.29**	0.77	0.53	0.63	**62.49**
	Sigmoid SVM	0.56	0.58	0.57	**55.86**	0.65	0.56	0.6	**55.24**	0.65	0.57	0.61	**55.68**
	Gaussian SVM	0.57	0.6	0.58	**57.66**	0.71	0.71	0.71	**64.81**	0.71	0.71	0.71	**64.65**
	Random Forest	0.59	0.62	0.6	**59.26**	0.71	0.72	0.71	**66.01**	0.71	0.72	0.71	**65.77**
	k-NN	0.59	1	0.74	**65.52**	0.67	0.95	0.79	**69.5**	0.67	0.94	0.78	**68.65**

6.5.2 Performance of the Proposed Defense Method

The proposed defense method is trained and tested on two real-world datasets. Its performance is demonstrated with various types of classifiers. The overall statistics of the proposed defense method on the CelebA dataset are presented in Table 6.1. The results show that before employing image restoration, the classifiers Linear SVM, Polynomial SVM, Sigmoid SVM, Gaussian SVM, RF and k-NN classifier detect with an accuracy of 98.75%, 98.5%, 72%, 98%, 98.5% and 97.5% respectively. Among the classifiers, Linear SVM shows its effectiveness in detecting facial adversarial images from the original with the highest accuracy of 98.75%. After employing image restoration such as BL followed by SR (BL+SR) to the adversarial images, the classification accuracy improves from 98.75% to 99% for

Linear SVM, from 72% to 93% for Sigmoid SVM, from 98% to 99% for Gaussian SVM, from 98.5% to 99% for RF and from 97.5% to 98% for k-NN on CelebA dataset with P-FGSM adversarial attack.

On the FFHQ dataset with both adversarial attacks StyleGAN and FLM, our method achieves $5 - 10\%$ improvement in the classification accuracy in almost all classification models, even if it achieves low classification accuracy before applying image restoration. The results on the FFHQ dataset before and after applying image restoration (BL+SR) with adversarial attacks StyleGAN and FLM are shown in Table 6.2. Our experimental results show that BL alone is sufficient sometimes to bring back the adversarial images into the original space, leading the classifier toward correct prediction. Thus, the results show that significant improvement in the detection accuracy after employing the image restoration BL+SR on the adversarial images.

6.6 Conclusions

In this chapter, we investigate a new defense model for improving robustness against facial adversarial attacks based on deep image restoration networks. We generate a well-protected version of adversarial face images based on P-FGSM, FLM and StyleGAN and have proved that these images can mislead the classifier to misclassification with high confidence. Image restorations such as BL followed by SR are performed on adversarial images to enhance the visual quality of images, which brings back the low-resolution adversarial images into the original high-resolution space. The encoded features are extracted for the recovered images and trained on various types of classifiers. The results are demonstrated on two real-world datasets for different types of adversarial attacks. The experimental results show that there is a significant improvement in the classification accuracy after employing image restoration in the classification models.

Chapter 7

Conclusions and Future Directions

The work in this thesis addressed the research problems of detecting digital image attacks. We proposed models for detecting and mitigating the effect of digital image attacks to protect CV-based models from misclassification.

First, we addressed the problem of detecting swapped face images from the original. We extracted augmented 81-facial landmarks which include facial landmarks on the forehead as well. A full face swapping is performed based on the extracted augmented 81-facial landmarks of both the source face and destination face image. We extracted encoded WLMP features from images and provided them to different types of SVMs for the detection of the presence of an attack. Numerical results show that Linear-SVM and Polynomial-SVM achieve a precision of 96% and a recall value of 94% for swapped face images with a detection accuracy of 95%. Whereas both Sigmoid-SVM and Gaussian-SVM achieve the same values for precision and recall for both fake and original images with an accuracy of 74%.

Next, we proposed a method for the localization and detection of CMF attacks in digital images. The objective is to accurately localize and detect CMFs present in the image. We extract SURF and BRISK descriptors. We match both fused features and perform clustering using the HAC method to reduce false positives.

We evaluate our detection method on real-world CMF datasets and experimental results are presented in terms of TPR and FPR with varying threshold t. We also compare the results of our detection method with the SOTA methods. Our detection method achieves the highest 98% TPR and the lowest 7.5% FPR at the threshold of 0.09. It outperforms some of the SOTA methods in terms of TPR and running time.

Next, we proposed two defense methods against facial adversarial attacks. The first method is against intensity-based adversarial attacks and another is against geometry-based facial adversarial attacks. We generate different kinds of facial adversarial attacks and also present the results obtained after facial adversarial attacks. Distinct feature descriptors are extracted from face images and provided to various types of classification models to defend against facial adversarial attacks from clean images. We evaluate our defense methods on real-world datasets. The performance of our methods is demonstrated with different types of classifiers. Numerical results show that our defense methods achieve significant performance on a wide range of adversarial attacks.

Finally, we proposed a model for improving facial adversarial robustness. We generate facial adversarial attacks based on different kinds of adversarial methods. We restore the facial adversarial images using image restoration techniques. That is, we bring back images into the original space from the adversarial space by applying BL filtering and image SR. WLMP features are extracted and fed into various classifiers. We evaluate our defense method on real-world datasets. We also present experimental results before and after image restoration techniques. The performance of our method is demonstrated on different kinds of classifiers. Numerical results show that the proposed model improves the adversarial robustness against a spectrum of adversarial attacks for different types of classification models.

7.1 Future Directions

Although the contributions of this thesis detect and mitigate the effect of digital image attacks on CV-based systems, constant defensive strategy improvement is necessary due to the dynamic nature of attacks. For instance, in the age of DL, the components of DL systems, such as filters, whole layers, or decision functions can also be attacked [128]. In addition, the creation of high-resolution humans, such as face photographs or generic object images is incredibly simple because of advancements in the development of synthetic images using generative networks. Such synthetic crafted images can be added to the training data through backdoor data poisoning [129]. Thus, there are several possible directions in which the work in this thesis can be extended. We list a few immediate extensions of our work.

- Since DL often requires large datasets and takes more time for training, cloud service providers such as Azure, Google, Baidu, AWS and Alibaba offer Applicant Program Interfaces (APIs) for their clients to accomplish CV tasks. Such APIs can help the users of cloud services to check images for both non-commercial and commercial purposes. Therefore, potential defense methods are required to protect vision APIs.

- It is observed that images contain a natural and intrinsic structure that can be leveraged to reverse many types of adversarial attacks. Extracting various kinds of intrinsic features of an image may play a critical role in reversing adversarial attacks.

- Facial adversarial attacks mainly focus on facial features to attack. The geometry of the face is a unique identity and provides distinct information about the face. Thus, geometry-based features such as facial landmarks, face embedding, etc., could improve adversarial robustness.

- A large-scale dataset with difficult attacks using advanced latex and silicone masks, image synthesis and morphing technologies in many imaging spectrums would help to detect image attacks effectively.

- CV-based models usually contain blocks such as feature extraction, matching and network manipulation. Recently, it has been demonstrated that traditional blocks of any deep learning model may be transformed into secure systems using blockchain technology. The security of CV-based models can be enhanced from the attacks with a combination of blockchain and ML framework [130].

References

[1] D. Deb, X. Liu, and A. K. Jain, "Unified detection of digital and physical face attacks," *arXiv preprint arXiv:2104.02156*, 2021.

[2] I. Amerini, L. Ballan, R. Caldelli, A. Del Bimbo, and G. Serra, "A sift-based forensic method for copy–move attack detection and transformation recovery," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 3, pp. 1099–1110, 2011.

[3] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.

[4] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8110–8119.

[5] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Learning rich features for image manipulation detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1053–1061.

[6] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceed-*

ings of the IEEE/CVF international conference on computer vision, 2019, pp. 1–11.

[7] J. Zhao and J. Guo, "Passive forensics for copy-move image forgery using a method based on dct and svd," *Forensic science international*, vol. 233, no. 1-3, pp. 158–166, 2013.

[8] H. A. Alberry, A. A. Hegazy, and G. I. Salama, "A fast sift based method for copy move forgery detection," *Future Computing and Informatics Journal*, vol. 3, no. 2, pp. 159–165, 2018.

[9] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1867–1874, 2014.

[10] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[11] A. Dabouei, S. Soleymani, J. Dawson, and N. Nasrabadi, "Fast geometrically-perturbed adversarial faces," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1979–1988.

[12] C. Xie, Y. Wu, L. v. d. Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 501–509.

[13] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," 2014.

[14] M. A. Oikawa, Z. Dias, A. de Rezende Rocha, and S. Goldenstein, "Manifold learning and spectral clustering for image phylogeny forests," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 1, pp. 5–18, 2015.

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

[16] J. Kietzmann, L. W. Lee, I. P. McCarthy, and T. C. Kietzmann, "Deepfakes: Trick or treat?" *Business Horizons*, vol. 63, no. 2, pp. 135–146, 2020.

[17] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 503–510.

[18] S. Mosaddegh, L. Simon, and F. Jurie, "Photorealistic face de-identification by aggregating donors face components," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 159–174.

[19] L. Wolf, Z. Freund, and S. Avidan, "An eye for an eye: A single camera gaze-replacement method," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 817–824.

[20] Y. Zhang, L. Zheng, and V. L. Thing, "Automated face swapping and its detection," in *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*. IEEE, 2017, pp. 15–19.

[21] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[22] A. J. Fridrich, B. D. Soukal, and A. J. Lukáš, "Detection of copy-move forgery in digital images," in *in Proceedings of Digital Forensic Research Workshop*. Citeseer, 2003.

[23] V. Christlein, C. Riess, J. Jordan, C. Riess, and E. Angelopoulou, "An evaluation of popular copy-move forgery detection approaches," *IEEE Transactions on information forensics and security*, vol. 7, no. 6, pp. 1841–1854, 2012.

[24] S. Bayram, I. Avcibas, B. Sankur, and N. Memon, "Image manipulation detection with binary similarity measures," in *Signal Processing Conference, 2005 13th European*. IEEE, 2005, pp. 1–4.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[26] L. Deng and Y. Liu, *Deep learning in natural language processing*. Springer, 2018.

[27] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[28] A. Ortis, G. M. Farinella, and S. Battiato, "An overview on image sentiment analysis: Methods, datasets and current challenges." in *ICETE (1)*, 2019, pp. 296–306.

[29] F. Carrara, A. Esuli, T. Fagni, F. Falchi, and A. M. Fernández, "Picture it in your mind: Generating high level visual representations from textual descriptions," *Information Retrieval Journal*, vol. 21, no. 2, pp. 208–229, 2018.

[30] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[31] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel, "On the (statistical) detection of adversarial examples," *arXiv preprint arXiv:1702.06280*, 2017.

[32] H. Etienne, "The future of online trust (and why deepfake is advancing it)," *AI and Ethics*, vol. 1, no. 4, pp. 553–562, 2021.

[33] L. Verdoliva, "Media forensics and deepfakes: an overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[34] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, 1997, pp. 353–360.

[35] 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.

[36] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.

[37] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, "Forensictransfer: Weakly-supervised domain adaptation for forgery detection," *arXiv preprint arXiv:1812.02510*, 2018.

[38] H. Huang, W. Guo, and Y. Zhang, "Detection of copy-move forgery in digital images using sift algorithm," in *2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application*, vol. 2, Dec 2008, pp. 272–276.

[39] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[40] B. Schneier, "Inside risks: The uses and abuses of biometrics," *Commun. ACM*, vol. 42, no. 8, p. 136, aug 1999. [Online]. Available: https://doi.org/10.1145/310930.310988

[41] M. Singh, R. Singh, and A. Ross, "A comprehensive overview of biometric fusion," *Information Fusion*, vol. 52, pp. 187–205, 2019.

[42] 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.

[43] U. Scherhag, R. Raghavendra, K. B. Raja, M. Gomez-Barrero, C. Rathgeb, and C. Busch, "On the vulnerability of face recognition systems towards morphed face attacks," in *2017 5th international workshop on biometrics and forensics (IWBF)*. IEEE, 2017, pp. 1–6.

[44] Y. Kim, J.-H. Yoo, and K. Choi, "A motion and similarity-based fake detection method for biometric face recognition systems," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 2, pp. 756–762, 2011.

[45] K. Kotwal, Z. Mostaani, and S. Marcel, "Detection of age-induced makeup attacks on face recognition systems using multi-layer deep features," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 2, no. 1, pp. 15–25, 2019.

[46] S. Bhattacharjee, A. Mohammadi, and S. Marcel, "Spoofing deep face recognition with custom silicone masks," in *2018 IEEE 9th international*

conference on biometrics theory, applications and systems (BTAS). IEEE, 2018, pp. 1–7.

[47] J. Hu, X. Liao, W. Wang, and Z. Qin, "Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1089–1102, 2021.

[48] V. Blanz, K. Scherbaum, T. Vetter, and H.-P. Seidel, "Exchanging faces in images," in *Computer Graphics Forum*, vol. 23, no. 3. Wiley Online Library, 2004, pp. 669–676.

[49] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar, "Face swapping: automatically replacing faces in photographs," in *ACM SIGGRAPH 2008 papers*, 2008, pp. 1–8.

[50] S. Mahajan, L.-J. Chen, and T.-C. Tsai, "Swapitup: A face swap application for privacy protection," in *2017 IEEE 31st International Conference on Advanced Information Networking and Applications (AINA)*. IEEE, 2017, pp. 46–50.

[51] 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.

[52] D. Chen, Q. Chen, J. Wu, X. Yu, and T. Jia, "Face swapping: realistic image synthesis based on facial landmarks alignment," *Mathematical Problems in Engineering*, vol. 2019, 2019.

[53] R. Natsume, T. Yatagawa, and S. Morishima, "Fsnet: An identity-aware generative model for image-based face swapping," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 117–132.

[54] ——, "Rsgan: face swapping and editing using face and hair representation in latent spaces," *arXiv preprint arXiv:1804.03447*, 2018.

[55] Y. Nirkin, Y. Keller, and T. Hassner, "Fsgan: Subject agnostic face swapping and reenactment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7184–7193.

[56] A. Agarwal, R. Singh, M. Vatsa, and A. Noore, "Swapped! digital face presentation attack detection via weighted local magnitude pattern," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 659–665.

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

[58] A. Khodabakhsh, R. Ramachandra, K. Raja, P. Wasnik, and C. Busch, "Fake face detection methods: Can they be generalized?" in *2018 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 2018, pp. 1–6.

[59] X. Ding, Z. Raziei, E. C. Larson, E. V. Olinick, P. Krueger, and M. Hahsler, "Swapped face detection using deep learning and subjective assessment," *EURASIP Journal on Information Security*, vol. 2020, pp. 1–12, 2020.

[60] V. Christlein, C. Riess, J. Jordan, C. Riess, and E. Angelopoulou, "An evaluation of popular copy-move forgery detection approaches," *arXiv preprint arXiv:1208.3665*, 2012.

[61] A. J. Fridrich, B. D. Soukal, and A. J. Lukáš, "Detection of copy-move forgery in digital images," in *in Proceedings of Digital Forensic Research Workshop*. Citeseer, 2003.

[62] A. Popescu and H. Farid, "Exposing digital forgeries by detecting duplicated image regions. department computer science, dartmouth college, technology report tr2004-515," 2004.

[63] S. Bayram, H. T. Sencar, and N. Memon, "An efficient and robust method for detecting copy-move forgery," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on.* IEEE, 2009, pp. 1053–1056.

[64] B. Mahdian and S. Saic, "Detection of copy–move forgery using a method based on blur moment invariants," *Forensic science international*, vol. 171, no. 2-3, pp. 180–189, 2007.

[65] J. Wang, G. Liu, Z. Zhang, Y. Dai, and Z. Wang, "Fast and robust forensics for image region-duplication forgery," *Acta Automatica Sinica*, vol. 35, no. 12, pp. 1488–1495, 2009.

[66] S.-J. Ryu, M.-J. Lee, and H.-K. Lee, "Detection of copy-rotate-move forgery using zernike moments," in *International Workshop on Information Hiding.* Springer, 2010, pp. 51–65.

[67] G. Muhammad, M. Hussain, and G. Bebis, "Passive copy move image forgery detection using undecimated dyadic wavelet transform," *Digital investigation*, vol. 9, no. 1, pp. 49–57, 2012.

[68] D. Cozzolino, G. Poggi, and L. Verdoliva, "Efficient dense-field copy–move forgery detection," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 11, pp. 2284–2297, 2015.

[69] B. Soni, P. K. Das, and D. M. Thounaojam, "Copy-move tampering detection based on local binary pattern histogram fourier feature," in *Proceedings of the*

7th International Conference on Computer and Communication Technology. ACM, 2017, pp. 78–83.

[70] H. Huang, W. Guo, and Y. Zhang, "Detection of copy-move forgery in digital images using sift algorithm," in *2008 IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application*, vol. 2. IEEE, 2008, pp. 272–276.

[71] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[72] X. Pan and S. Lyu, "Region duplication detection using image feature matching," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 4, pp. 857–867, 2010.

[73] E. Ardizzone and G. Mazzola, "Detecting multiple copies in tampered images," in *ICIP 2010-17th IEEE International Conference on Image Processing*, 2010.

[74] B. Shivakumar and S. S. Baboo, "Detection of region duplication forgery in digital images using surf," *International Journal of Computer Science Issues (IJCSI)*, vol. 8, no. 4, p. 199, 2011.

[75] J. Zhao and W. Zhao, "Passive forensics for region duplication image forgery based on harris feature points and local binary patterns," *Mathematical Problems in Engineering*, vol. 2013, 2013.

[76] B. Soni, P. K. Das, and D. M. Thounaojam, "multicmfd: fast and efficient system for multiple copy-move forgeries detection in image," in *Proceedings of the 2018 International Conference on Image and Graphics Processing*. ACM, 2018, pp. 53–58.

[77] P. Mishra, N. Mishra, S. Sharma, and R. Patel, "Region duplication forgery detection technique based on surf and hac," *The Scientific World Journal*, vol. 2013, 2013.

[78] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2013, pp. 387–402.

[79] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical programming*, vol. 45, no. 1, pp. 503–528, 1989.

[80] A. Kurakin, I. Goodfellow, S. Bengio *et al.*, "Adversarial examples in the physical world," 2016.

[81] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.

[82] A. Rozsa, E. M. Rudd, and T. E. Boult, "Adversarial diversity and hard positive generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 25–32.

[83] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 2016, pp. 372–387.

[84] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2574–2582.

[85] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, pp. 506–519.

[86] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 15–26.

[87] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2137–2146.

[88] M. Alzantot, Y. Sharma, S. Chakraborty, H. Zhang, C.-J. Hsieh, and M. B. Srivastava, "Genattack: Practical black-box attacks with gradient-free optimization," in *Proceedings of the Genetic and Evolutionary Computation Conference*, 2019, pp. 1111–1119.

[89] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," *arXiv preprint arXiv:1801.02610*, 2018.

[90] 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.

[91] D. Deb, J. Zhang, and A. K. Jain, "Advfaces: Adversarial face synthesis," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2020, pp. 1–10.

[92] H. Qiu, C. Xiao, L. Yang, X. Yan, H. Lee, and B. Li, "Semanticadv: Generating adversarial examples via attribute-conditioned image editing," in *European Conference on Computer Vision*. Springer, 2020, pp. 19–37.

[93] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song, "Spatially transformed adversarial examples," *arXiv preprint arXiv:1801.02612*, 2018.

[94] D. Hendrycks and K. Gimpel, "Early methods for detecting adversarial images," *arXiv preprint arXiv:1608.00530*, 2016.

[95] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[96] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," *arXiv preprint arXiv:1705.07204*, 2017.

[97] Z. Gong, W. Wang, and W.-S. Ku, "Adversarial and clean data are not twins," *arXiv preprint arXiv:1704.04960*, 2017.

[98] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations," *arXiv preprint arXiv:1702.04267*, 2017.

[99] F. V. Massoli, F. Carrara, G. Amato, and F. Falchi, "Detection of face recognition adversarial attacks," *Computer Vision and Image Understanding*, vol. 202, p. 103103, 2021.

[100] A. Agarwal, R. Singh, M. Vatsa, and N. Ratha, "Are image-agnostic universal adversarial perturbations for face recognition difficult to detect?" in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2018, pp. 1–7.

[101] A. Mustafa, S. H. Khan, M. Hayat, J. Shen, and L. Shao, "Image super-resolution as a defense against adversarial attacks," *IEEE Transactions on Image Processing*, vol. 29, pp. 1711–1724, 2019.

[102] P. Huber, G. Hu, R. Tena, P. Mortazavian, P. Koppen, W. J. Christmas, M. Ratsch, and J. Kittler, "A multiresolution 3d morphable face model and fitting framework," in *Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016.

[103] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[104] C. Cao, Y. Weng, S. Lin, and K. Zhou, "3d shape regression for real-time facial animation," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 4, p. 41, 2013.

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

[106] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2387–2395.

[107] J. Lv, X. Shao, J. Xing, C. Cheng, and X. Zhou, "A deep regression architecture with two-stage re-initialization for high performance facial landmark detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3317–3326.

[108] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 532–539.

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

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

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

[112] S. Leutenegger, M. Chli, and R. Y. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *2011 International conference on computer vision*. Ieee, 2011, pp. 2548–2555.

[113] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.

[114] A. J. Fridrich, B. D. Soukal, and A. J. Lukáš, "Detection of copy-move forgery in digital images," in *in Proceedings of Digital Forensic Research Workshop*. Citeseer, 2003.

[115] C. Y. Li, A. S. Shamsabadi, R. Sanchez-Matilla, R. Mazzon, and A. Cavallaro, "Scene privacy protection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2502–2506.

[116] A. Kurakin, I. Goodfellow, S. Bengio, Y. Dong, F. Liao, M. Liang, T. Pang, J. Zhu, X. Hu, C. Xie *et al.*, "Adversarial attacks and defences competition," in *The NIPS'17 Competition: Building Intelligent Systems*. Springer, 2018, pp. 195–231.

[117] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," *arXiv preprint arXiv:1704.01155*, 2017.

[118] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, L. Chen, M. E. Kounavis, and D. H. Chau, "Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression," *arXiv preprint arXiv:1705.02900*, 2017.

[119] N. Krawetz and H. F. Solutions, "A pictures worth," *Hacker Factor Solutions*, vol. 6, no. 2, p. 2, 2007.

[120] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," *arXiv preprint arXiv:1412.5068*, 2014.

[121] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE symposium on security and privacy (SP)*. IEEE, 2016, pp. 582–597.

[122] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017, pp. 3–14.

[123] C. Sadu and P. K. Das, "A defense method against facial adversarial attacks," in *TENCON 2021-2021 IEEE Region 10 Conference (TENCON)*. IEEE, 2021, pp. 459–463.

[124] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*. IEEE, 1998, pp. 839–846.

[125] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1646–1654.

[126] ——, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1637–1645.

[127] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[128] S. Sabour, Y. Cao, F. Faghri, and D. J. Fleet, "Adversarial manipulation of deep representations," *arXiv preprint arXiv:1511.05122*, 2015.

[129] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[130] A. Goel, A. Agarwal, M. Vatsa, R. Singh, and N. Ratha, "Deepring: Protecting deep neural network with blockchain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Publications Related to Thesis

Journals

1. **Chiranjeevi Sadu** and Pradip K. Das, "Image Restoration Networks for Improving Facial Adversarial Robustness", Journal of Visual Communication and Image Representation. [**Under Review**] [Chapter 6]

Book Chapter

1. **Chiranjeevi Sadu** and Pradip K. Das, "Detection of Augmented Facial Landmarks-based Face Swapping", Computer Vision Applications of Visual AI and Image Processing (CVIP) 2021, De Gruyter. [**Accepted**] [Chapter 3]

Conference Proceedings

1. **Chiranjeevi Sadu** and Pradip K. Das, "Swapping face images based on augmented facial landmarks and its detection.", In 2020 IEEE REGION 10 CONFERENCE (TENCON), pp. 456-461. IEEE, 2020. [Chapter 3]

2. **Chiranjeevi Sadu** and Pradip K. Das, "A Detection Method for Copy-move Forgery Attacks in Digital Images", IEEE REGION 10 CONFERENCE (TENCON) 2022. [**Presented**] [Chapter 4]

3. **Chiranjeevi Sadu** and Pradip K. Das, "A Defense Method Against Facial Adversarial Attacks", In TENCON 2021-2021 IEEE Region 10 Conference (TENCON), pp. 459-463. IEEE, 2021. [Chapter 5]

4. **Chiranjeevi Sadu** and Pradip K. Das, "Detection of Geometry-based Adversarial Facial Attacks using Error Level Analysis", IEEE REGION 10 CONFERENCE (TENCON) 2022. [**Presented**] [Chapter 5]

Publications Outside Thesis

1. Singh, Krishna Kumar, Amit Patel and **Chiranjeevi Sadu**, "Correlation Scaled Principal Component Regression" In International Conference on Intelligent Systems Design and Applications, pp. 350-356. Springer, Cham, 2017.

2. Singh, Krishna Kumar, **Sadu Chiranjeevi** and Kethavath Sivalal, "Face Features-based Personality Assessment", International Conferences Computer Graphics, Visualization, Computer Vision and Image Processing, pp. 45-52, 2021.