DEEP LEARNING FRAMEWORKS FOR FACE ANTI-SPOOFING

Thesis Submitted
in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in

Electronics and Communication Engineering

by

AASHANIA ANTIL

(Enrollment No.: 2K20/PHDEC/504)

Under the supervision of

DR. CHHAVI DHIMAN
Assistant Professor in ECE Department

To the

Department of Electronics and Communication Engineering

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

December, 2025

ACKNOWLEDGEMENTS

The completion of this doctoral journey has been an extraordinary experience, and I am deeply grateful to all those who have contributed to its success. Their support, guidance, and encouragement have been invaluable.

First and foremost, I express my sincere gratitude to the Almighty for granting me the strength, resilience, and determination to pursue this endeavor. His blessings have been a constant source of inspiration throughout this journey.

I am profoundly thankful to my research supervisor, **Dr. Chhavi Dhiman**, Department of Electronics and Communication Engineering, Delhi Technological University. Her expert guidance, insightful feedback, and unwavering support have been instrumental in shaping this work. Her mentorship has been a privilege that has greatly influenced my academic growth.

I extend my heartfelt thanks to my friends— **Mr. Sachin, Mr. Rohit, Ms. Khyati,** and **Ms. Diksha**— for their encouragement and emotional support during challenging times. Their belief in me has been a source of strength throughout this process.

I also wish to acknowledge my colleagues in the Department of Electronics and Communication Engineering, particularly **Ms. Neha Sharma, Ms. Hemanshi,** and **Mr. Gaurav, Mr. Rohit** for their cooperation and valuable discussions that enriched my research experience.

My deepest gratitude goes to my family for their unwavering love and support. I am indebted to my parents, **Mr. Surender Singh**, and **Ms. Sunita**, as well as my siblings, **Ms. Yamini Antil**, and **Mr. Sagar Antil**, for their sacrifices and encouragement that fueled my determination to succeed. A special mention goes to my grandparents, **Late Ran Singh** and **Ms. Anguri Devi**, whose faith in me has been a guiding force throughout my life.

Finally, I am grateful to my husband, **Mr. Akash Rangi** and father-in-law, Mr. **Azad Singh** for their understanding and support during this journey. Their encouragement allowed me the time and focus needed to complete this work.

This thesis is a testament to the collective efforts of all these remarkable

individuals who have supported me along the way. I offer them my deepest thanks for their contributions to this milestone in my life.

AASHANIA ANTIL

DELHI TECHNOLOGICAL UNIVERSITY

Formerly Delhi College of Engineering

Shahbad Daulatpur, Main Bawana Road, Delhi –42

CANDIDATE’S DECLARATION

I **Aashania Antil** hereby certify that the work which is being presented in the thesis entitled **Deep Learning Frameworks for Face Anti-spoofing** in partial fulfillment of the requirements for the award of the Degree of Doctor in Philosophy, submitted in the **Department of Electronics and Communication Engineering**, Delhi Technological University is an authentic record of my own work carried out during the period from **January 2021** to **December 2025** under the supervision of **Dr. Chhavi Dhiman**.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Candidate’s Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Dr. Chhavi Dhiman
Supervisor
Assistant Professor, ECE Dept.
DTU, Delhi

Prof. Sumantra Dutta Roy
External Examiner
Professor, EE Dept.
IIT Delhi

DELHI TECHNOLOGICAL UNIVERSITY

Formerly Delhi College of Engineering

Shahbad Daulatpur, Main Bawana Road, Delhi –42

CERTIFICATE BY THE SUPERVISOR(S)

Certified that **Aashania Antil** (Enrollment No.: 2K20/PHDEC/504) has carried out their research work presented in this thesis entitled "**Deep Learning Frameworks for Face Anti-spoofing**", for the award of **Doctor of Philosophy** from the Department of Electronics and Communication Engineering, Delhi Technological University, under our guidance and supervision. The thesis embodies results of original work, and studies are carried out by the student herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Dr. Chhavi Dhiman
Supervisor
Department of ECE
Delhi Technological University,
Delhi –110042, India,

Place and Date:

DEEP LEARNING FRAMEWORKS FOR FACE ANTI-SPOOFING

AASHANIA ANTIL

ABSTRACT

With the rapid integration of facial recognition (FR) systems in access control, banking, and mobile authentication, the risk of face spoofing—also known as presentation attacks (PAs)—has grown significantly. These attacks, carried out using printed images, replayed videos, or 3D masks, threaten the security and reliability of biometric authentication systems. The increasing sophistication of spoofing techniques, combined with the low cost and easy availability of generative tools, underscores the urgent need for robust Face Anti-Spoofing (FAS) or Presentation Attack Detection (PAD) mechanisms. Although several approaches have been proposed, many existing solutions struggle to generalize under challenging conditions involving varied lighting, spoofing materials, backgrounds, and image/video quality.

To address these challenges, this thesis proposes a suite of deep learning-based frameworks that are robust, interpretable, and generalizable for real-world face anti-spoofing. The research is structured around four complementary solutions, each targeting a specific dimension of the problem: texture-based learning, multi-modal fusion, spatio-temporal modeling, and generative learning.

The first solution introduces a two-stream hybrid framework that fuses handcrafted and deep features to improve spoof detection accuracy. It combines Multi-Level Extended Local Binary Patterns (ELBP) to capture fine-grained texture information with a modified Xception network, enhanced by Squeeze-and-Excitation (SE) blocks for channel-wise feature reweighting without increasing complexity. This design balances expressive power and computational efficiency, enabling the model to handle diverse spoofing conditions and maintain generalization across datasets.

The second solution, MF^{2}ShrT, addresses multi-modal fusion by leveraging the power of Vision Transformers (ViTs). It uses overlapping patches to emphasize local contextual cues and introduces SharLViT, a shared-layer transformer backbone that improves feature representation while reducing parameter count. A novel T-Encoder-based Hybrid Feature Block is employed to mine inter-modal dependencies across RGB, depth, and IR streams. The Adaptive Weighted Fusion and Classification Block (AWFCB) then learns to dynamically combine these features, emphasizing salient cues while suppressing redundant information—resulting in a flexible and accurate spoof detection system.

The third solution focuses on the temporal dimension of FAS by proposing Bi-STAM, a Bi-Directional Spatio-Temporal Adaptive Modeling framework. Aiming to capture motion inconsistencies and subtle dynamics in video-based attacks, it introduces two key components: a Temporal Adaptive Block (TAB) to balance motion and static information, and a Spatial Adaptive Block (SAB) to enhance texture representation while filtering noise. These are fused via a Feature Aggregation Block (FAB) to yield a unified spatio-temporal representation, significantly boosting generalization and performance on video-based spoof detection tasks.

The fourth solution, PolarSentinelGAN, presents a novel generative adversarial framework that enhances spoof classification through depth map generation. By fusing RGB and Multi-Scale Retinex with Color Preservation (MSRCP) inputs, the model uses Dual Polarized Attention (DPAttn) to focus on discriminative regions. A dedicated Feed Forward Block (FFB) within the generator facilitates the transmission of rich features, while optimized latent variables improve generalization across attack types and datasets.

All four frameworks are extensively evaluated using standard intra- and cross-dataset testing protocols on public benchmarks, and are further supported with explainability techniques such as class activation mapping and feature occlusion testing. The results demonstrate strong real-time performance, robustness, and scalability.

The proposed methodologies in this thesis makes substantial contributions toward the development of next-generation face anti-spoofing systems. The proposed methods not only address key challenges in generalization, efficiency, and interpretability, but also pave the way for practical deployment in critical domains such as finance, border security, surveillance, and consumer electronics. Future directions include exploring

federated learning, privacy-aware architectures, and continual domain adaptation to further enhance system reliability in dynamic real-world environments.

LIST OF PUBLICATIONS

Journal Papers

- **A. Antil** and C. Dhiman "Unmasking Deception: A Comprehensive Survey on the Evolution of Face Anti-spoofing Methods," *Neurocomputing*, vol. 617, 2025. SCIE (IF: 5.5). DOI: https://doi.org/10.1016/j.neucom.2024.128992 (accepted and published)
- **A. Antil** and C. Dhiman, "A two stream face anti spoofing framework using multi-level deep features and ELBP features," *Multimedia Systems*, 2023. SCIE (IF: 3.5). DOI: https://doi.org/10.1007/s00530-023-01060-7 (accepted and published)
- **A. Antil** and C. Dhiman, "MF^{2}ShrT: Multimodal Feature Fusion Using Shared Layered Transformer for Face Anti-spoofing," *ACM Transactions on Multimedia Computing, Communications, and Applications,* vol. 20, no. 6, pp. 1-21, 2024. SCIE (IF:5.2). DOI: https://doi.org/10.1145/3640817 (accepted and published)
- C. Dhiman, **A. Antil**, A. Anand, and S. Gakhar, "A deep face spoof detection framework using multi-level ELBPs and stacked LSTMs," *Signal, Image and Video Processing*, vol. 18, p. 499–512, 2024. SCIE (IF:2.0). DOI: https://doi.org/10.1007/s11760-024-03169-2 (accepted and published)

Conference Papers

- **A. Antil** and C. Dhiman, "Two Stream RGB-LBP Based Transfer Learning Model for Face Anti-spoofing," in *7th International Conference on Computer Vision & Image Processing (CVIP)*, India, 2023. DOI: https://doi.org/10.1007/978-3-031-31407-0_28 (accepted and published)
- **A. Antil** and C. Dhiman, "Securing Faces: A GAN-Powered Defense Against Spoofing with MSRCR and CBAM," in 27th International Conference on Pattern Recognition (ICPR), Kolkata, 2024. DOI: https://doi.org/10.1007/978-3-031-78201-5_28 (accepted and published)
- **A. Antil** and C. Dhiman, "Leveraging Depth Data and Parameter Sharing in Vision

Transformers for Improved Face Anti-Spoofing," *6th International Conference on Artificial Intelligence and Speech Technology (AIST)***,** IGDTUW, Delhi, 2024. DOI:https://link.springer.com/chapter/10.1007/978-3-031-91340-2_13 (accepted and published)

Communicated

- **A. Antil** and C. Dhiman, "PolarSentinelGAN: A Dual-Polarized Attention-Guided Generative Adversarial Framework for Robust Face Anti-Spoofing," *IEEE Transactions on Dependable and Secure Computing (TDSC).* SCIE (IF:7.5). (Under Review)
- **A. Antil** and C. Dhiman, "Bi-STAM: Bi-Directional Spatio-Temporal Adaptive Modeling for Robust Face Anti-Spoofing," *Knowledge-Based Systems*. SCIE (IF:7.6). (Under Review)

TABLE OF CONTENTS

Acknowledgment iii-iv

Candidate’s Declaration v

Certificate by the Supervisor(s) vi

Abstract vii-ix

List of Publications x-xi

Table of Contents xii-xv

List of Tables xvi-xviii

List of Figures xix-xxiii

List of Abbreviations xxiv-xxvi

Chapter 1: Introduction 1-16

1.1 Face Spoofing and Anti-spoofing: An Overview

1.1.1 Human Characteristics

1.1.2 Non-Human Characteristics

1.1.3 Hardware-based Methods

1.1.4 Software-based Methods

1.2 Challenges in the Field of Face Anti-spoofing

1.3 Research Motivation

1.4 Problem Formulation

1.5 Research Objectives

1.6 Research Contribution

1.7 Significance of the Study

1.8 Outlines of the Thesis

Chapter 2: Literature Review 17-38

2.1 Fundamental Key Features for FAS

2.1.1 Texture Methods

2.1.2 Recapturing-based Methods

2.1.3 Image Quality Analysis

2.1.4 Liveness Detection

2.1.5 Hybrid-based Methods

2.1.6 Discussion, Challenges and Limitations
2.2 FAS Design Approaches
2.2.1 Direct Supervision based Methods
2.2.2 3D Geometric based Methods
2.2.3 Unseen Domain based Methods
2.2.4 Unknown Attacks based Methods
2.2.5 Multi-Spectral based Methods
2.2.6 Composite Methods
2.3 FAS Backbone Architectures
2.4 Research Gaps
Chapter 3: Two-stream RGB-based FAS framework 39-55
3.1 Methodology
3.1.1 Pre-processing step
3.1.2 Color space
3.1.3 Feature Extraction
3.1.3.1 ELBP based texture features
3.1.3.2 Modified Xception based deep features
3.1.4 Classification Model
3.2 Database and Experimental Analysis
3.2.1 Dataset
3.2.2 Evaluation Metrics
3.2.3 Implementation Details
3.2.4 Comparative Analysis with other State-of-the-Arts
3.2.4.1 Intra-dataset testing
3.2.4.2 Inter-dataset testing
3.2.5 Ablation Analysis
3.3 Significant Outcomes
Chapter 4: Multimodal Vision Transformer for Robust Face Anti-Spoofing 56-71
4.1 Methodology
4.1.1 Input Adaptor

4.1.2 SharLViT: Parameter sharing in ViT
4.1.3 T-Encoder-based Hybrid Feature Block
4.1.4 Adaptive Weighted Fusion and Classification Block
4.2 Database and Experimental Analysis
4.2.1 Datasets
4.2.2 Evaluation Metrics
4.2.3 Implementations Details
4.2.4 Comparative Analysis with other State-of-the-Arts
4.2.5 Ablation Study
4.3 Significance Outcomes

Chapter 5: Spatio-Temporal Adaptive Modeling for Face Anti-Spoofing 72-88

5.1 Methodology
5.1.1 Bi-directional Temporal Difference (BiD)
5.1.2 Spatio-Temporal Adaptive Modeling Block (STAM)
5.1.3 Feature Aggregator Block and Classification Block
5.2 Database and Experimental Analysis
5.2.1 Datasets
5.2.2 Evaluation Metrics
5.2.3 Implementations Details
5.2.4 Comparative Analysis with other State-of-the-Arts
5.2.5 Ablation Study
5.2.6 Analysis of Framework Complexity
5.2.7 Visualization
5.3 Significance Outcomes

Chapter 6: Generative Learning-Based Pixel-Wise Face Anti-Spoofing Framework 89-120

6.1 Methodology
6.1.1 Multi-scale Retinex with Color Restoration (MSRCR)

6.1.2 Convolution Block Attention Module (CBAM)

6.1.3 Network Architecture

6.2 Experimental Analysis

6.2.1 Implementation Details

6.2.2 Comparative Analysis with other State-of-the-Arts

6.2.3 Ablation Study

6.2.4 Visualization and Analysis

6.3 Significant Outcomes

6.4 Methodology

6.4.1 Multi-Scale Retinex with Color Preservation (MSRCP)

6.4.2 Dual Polarized Self-Attention Guided Module (DPAttn)

6.4.3 Generative Network

6.4.4 Classification

6.5 Experimental Analysis

6.5.1 Implementation Details

6.5.2 Comparative Analysis with other-State-of-the-Arts

6.5.3 Ablation Study

6.5.4 Visualization and Analysis

6.6 Significant Outcomes

Chapter 7: Conclusion, Future Scope, and Social Impact 121-125

7.1 Summary of the Work Done in the Thesis

7.2 Resource Efficiency of the Proposed FAS Frameworks

7.3 Future Scope/ Directions and Social Impact

References 126-143

List of Publications Related to the Thesis and their Proofs 144-151

Plagiarism Report 152-162

Curriculum Vitae/ Brief Profile 164-165

LIST OF TABLES

Table No.	Title	Page No.
Table 2.1	Comparative Overview of the Backbones employed in FAS.	37
Table 3.1	Intra-dataset testing: CASIA-FASD, Replay-Attack and MSU-MFSD datasets (%).	48
Table 3.2	Cross-dataset evaluation on MSU-MFSD, CASIA FASD, Replay-Attack and in terms of HTER (%).	49
Table 3.3	Ablation Study: Intra-dataset evaluation of modified Xception architecture on Reply-Attack, CASIA-FASD, and MSU-MFSD (%).	52
Table 3.4	Cross-dataset evaluation of Level-0 ELBP across different color spaces on CASIA-FASD, MSU-MFSD and Reply-Attack for ablation analysis (HTER %).	53
Table 3.5	Selection of Nodes in Dense Layers (D_i)	54
Table 4.1	Performance on the WMCA dataset during Intra-dataset Testing (%).	63
Table 4.2	Performance on the CASIA-SURF dataset during Intra-dataset Testing (%).	63
Table 4.3	Performance Evaluation for Cross-dataset Testing in terms of HTER (%).	64
Table 4.4	Ablation study of the effect of overlapping rate on the performance of WMCA dataset (%).	65
Table 4.5	Ablation study of baseline model and the effect of parameter sharing in the ViT model of WMCA dataset (%).	66
Table 4.6	Ablation study of effect of different Fusion techniques on the performance of WMCA dataset	67
Table 4.7	Comparison the performance of the WMCA	69

dataset using different losses (%).

Table 4.8 Comparison of computation efficiency for various models on the WMCA dataset. 69

Table 5.1 Evaluation results on CASIA-MFSD (C) dataset (%). 80

Table 5.2 Evaluation results on Replay-Attack (RA) dataset (%). 80

Table 5.3 Evaluation results on Oulu-NPU(O) dataset (%). 81

Table 5.4 Cross-Dataset Performance Evaluation: CASIA-FASD vs. Replay-Attack (HTER%). 81

Table 5.5 Cross-Dataset Performance Evaluation: OULU-NPU(O), MSU-MFSD (M), CASIA-MFSD (C) and REPLAY-ATTACK (RA) in HTER (%) and AUC (%). 82

Table 5.6 Performance evaluation of various network configurations on Protocol 2 of the OULU-NPU dataset, using 8 selected frames (%). 83

Table 5.7 Comparative Performance Analysis of Various Connection Strategies for TAB and SAB (%). 83

Table 6.1 Intra-Dataset evaluation on CASIA-FASD(C) Dataset (%). 96

Table 6.2 Intra-Dataset evaluation on Replay-Attack (RA) Dataset (%). 96

Table 6.3 Intra-Dataset evaluation on OULU-NPU (O) (%). 97

Table 6.4 Comparative Cross-Dataset Analysis: CASIA-FASD vs. Replay-Attack (HTER %). 98

Table 6.5 Comparative Cross-Dataset analysis across four datasets (%). 98

Table 6.6 Ablation Study Results: OULU-NPU Protocol 2 (%). 99

Table 6.7 Intra-dataset testing comparison within CASIA- 113

	FASD(C) (%).	
Table 6.8	Intra-dataset testing comparison Replay-Attack (RA) (%).	113
Table 6.9	Intra-dataset testing comparison ROSE-Youtu (%).	113
Table 6.10	Intra-dataset testing comparison OULU-NPU (O) (%).	114
Table 6.11	Comparative Cross-dataset analysis of CASIA-FASD, Replay-Attack and ROSE-Youtu (HTER %).	115
Table 6.12	Comparative Cross-Dataset analysis on four datasets (%).	116
Table 6.13	Ablation study on OULU-NPU protocol 2(%).	117
Table 7.1	Comparison of Proposed Frameworks: Resource Usage and Practical Applicability	123

LIST OF FIGURES

Figure No.	Title	Page No.
Figure 1.1	Publications over recent years in the field of Face Anti-Spoofing (FAS) obtained through Google Trends.	2
Figure 1.2	Direct and indirect attacks on FR security system [1]	2
Figure 1.3	Types of Face Attacks.	3
Figure 1.4	Evolution of Software based Face Anti-spoofing methods.	6
Figure 1.5	Overview of Current Face Spoofing Countermeasures.	7
Figure 1.6	Representative diagram to depict the problem formulation approach utilized in this thesis.	11
Figure 2.1	Classification of FAS Design Approaches.	20
Figure 2.2	Diverse Backbones employed in FAS domain.	33
Figure 2.3	Milestone of Face Anti-spoofing frameworks. The rise of handcrafted local descriptors in the early 2000s, followed by the introduction of traditional feature learning approaches in the late 2000s. In 2014, CNN achieved a breakthrough, prompting a shift to deep learning-based strategies such as 3D geometric and multi-spectral methods. The representation pipeline's deepening led to improved performance through Domain Generalization (DG) and Domain Adaptation (DA). Notably, within three years, Vision Transformer (ViT) and Generative Adversarial Network (GAN) integration significantly boosted FAS performance.	36
Figure 3.1	Proposed two-stream face anti-spoofing framework with multi-level ELBP and modified Xception.	40
Figure 3.2	Horizontal and Vertical ELBP Pattern.	42

Figure 3.3 An original image (a), $ELBP^{8,5,3}$ (b), $ELBP^{8,3,5}$ (c). 42

Figure 3.4 Multi-Level ELBP based texture feature extraction. 43

Figure 3.5 (a) Modified Xception Network; (b) Xception with integrated Squeeze and Excitation (SE) Block [2]. 44

Figure 3.6 ROC curves for (a) intra-dataset testing; (b) Inter-dataset testing: Test case 1- Trained on CASIA-FASD, Tested on Replay-Attack and MSU-MFSD, (c) Inter-dataset testing: Test Case 2- Trained on Replay-Attack, Tested on CASIA FASD and MSU-MFSD, (d) Inter-dataset testing: Test case 3- Trained on MSU-MFSD, Tested on CASIA- FASD and Replay-Attack. 51

Figure 3.7 Comparison of different generic architectures for baseline selection. 51

Figure 3.8 Intra-dataset testing EER and HTER obtained on (a) CASIA FASD (b) MSU-MFSD (c) Replay-Attack. 53

Figure 3.9 Performance of the proposed model with different designs on (a) CASIA-FASD (b) MSU-MFSD (c) Replay Attack using Intra-dataset testing. 54

Figure 4.1 The Proposed framework MF^{2}ShrT for Face Anti-spoofing 57

Figure 4.2 Illustration of Overlapping Patch Rate for different values. 58

Figure 4.3 SharLViT- Sequential Parameter Sharing in Vision Transformer Encoder Blocks. TE1, TE2, TE3 and TE4 denote 4 encoder blocks within a Transformer. Case 1: Without Parameter Sharing- Each encoder block has parameters P1, P2, P3 and P4 respectively. Case 2: With Parameter Sharing: P1 parameter is shared between 59 encoders TE1 and TE2, while the P2 parameter is shared between encoders TE3 and TE4.

Figure 4.4 Detailed structure of feature aggregator. 61

Figure 4.5 ROC curves for both Intra-dataset and Cross-dataset Testing on WMCA and CASIA-SURF. 65

Figure 4.6 Types Parameter Sharing Strategies (a) Sequence Sharing (b) Cyclic Sharing (c) Reverse Cyclic Sharing. 66

Figure 4.7 ROC curve for (a) parameter sharing methods, (b) Effect of weighted fusion classification (WFC) block. 67

Figure 4.8 Curve of the adaptively assigned weighted over epochs. 68

Figure 4.9 Trade-off between ACER and computational efficiency for various models under LOO protocol of WMCA dataset. 70

Figure 5.1 Illustration of the proposed Bi-STAM FAS framework. 73

Figure 5.2 Illustration of (a) TAB: Temporal Adaptive Block, (b) SAB: Spatial Adaptive Block- An example with an input of three frames (T=3). 75

Figure 5.3 Illustration of Feature Aggregator Block (FAB). 77

Figure 5.4 Adaptive weight progression over epochs: Initial weights $\{w_1, w_2, w_3\}$ are displayed, demonstrating their convergence to distinct values, indicating the respective contributions of $\mathcal{F}_1^s$, $\mathcal{F}_2^s$, and $\mathcal{F}_3^s$ 84

Figure 5.5 Visualization of ACER (%) on the OULU-NPU P1 dataset versus the number of parameters (M), highlighting the trade-off between model complexity and performance. 85

Figure 5.6 Visualization of Grad-CAM maps for the proposed Bi-STAM framework trained for 5, 15, and 55 epochs. Green-boxed samples represent live examples, while red-boxed samples indicate spoof examples. The first column presents the input samples, the second column shows the corresponding attention maps, and the third column displays the predictions for each sample. 86

Figure 5.7 t-SNE feature distributions for OULU-NPU Protocol 1 at various training stages: (a) Epoch 5, (b) Epoch 15, (c) Epoch 55. 86

Figure 6.1 Overview of the proposed GAN-based framework. 90

Figure 6.2 Pre-processes CASIA-FASD images: Green boundaries indicate live samples, red boundaries represent attack types. Top row shows RGB; Bottom row shows MSRCR. 91

Figure 6.3 Schematic diagram of CBAM Module [3]. 92

Figure 6.4 Comparison between Generated depth maps with Ground Truths maps. Row1: Input RGB images; Row2: Ground-Truth depth maps; and Row 3: Depth maps generated via proposed framework. The blue box highlights successful cases, while the red box displays failure cases. 100

Figure 6.5 t-SNE feature distribution for (a) CASIA-FASD (C), (b) REPLAY-ATTACK (RA). 100

Figure 6.6 An outline of the proposed PolarSentinelGAN Framework. It consists of three key components: a generator ($\mathbb{G}$) responsible for generating depth maps from RGB images, a discriminator ($\mathbb{D}$) tasked with appraising the quality of the generator's outputs, and a 103 classifier(ℂ) designed to distinguish genuine/spoofed face.

Figure 6.7 Pre-processed CASIA-FASD Images –green boundaries (Live samples) and red boundaries (different types of attacks), arranged from Top to Down: RGB, MSRCR, and MSRCP. 106

Figure 6.8 (a) Polarised Self-Channel Attention (PS-CA), (b) Polarised Self-Spatial Attention (PS-SA). 107

Figure 6.9 The Generator Architecture of our proposed framework PolarSentinelGAN. The generator in PolarSentinelGAN features a dual-stream design to enhance depth map reconstructions. The first stream, called Main Stream, 𝔾EnS utilizes encoder blocks to process RGB images as input. Simultaneously, the Attention- Reference Stream, 𝔾AttS operates as a second stream, employing a dual polarized self-attention guided attention module (DPAttn) to effectively process both RGB and MSRCP. 109

Figure 6.10 Comparison between generated depth images and Ground Truths for given RGB inputs. Row1: RGB domain input images; Row2: Ground-Truth depth maps; and Row 3: Depth maps generated. Green box-Columns 1 to 6 display success cases, while red box-columns 7 to 9 depict failure cases. 118

Figure 6.11 Visualizing Grad-CAM attention maps for various test samples, the first row demonstrates focused attention on live faces, while the second row showcases attention patterns learned for spoofed faces. 119

Figure 6.12 t-SNE feature distribution for (a) CASIA-FASD, (b) REPLAY-ATTACK. 119

LIST OF ABBREVIATIONS

FR	Facial Recognition
PAs	Presentation Attacks
PAD	Presentation Attack Detection
FAS	Face Anti-Spoofing
ELBP	Elliptical Local Binary Patterns
SE	Squeeze-and-Excitation
ViTs	Vision Transformers
AWFCB	Adaptive Weighted Fusion and Classification Block
TAB	Temporal Adaptive Block
SAB	Spatial Adaptive Block
FAB	Feature Aggregation Block
MSRCP	Multi-Scale Retinex with Color Preservation
DPAttn	Dual Polarized Attention
FFB	Feed Forward Block
LBP	Local Binary Patterns
HOG	Histogram of Oriented Gradients
SVM	Support Vector Machine
LDA	Linear Discriminant Analysis
LPQ	Local Phase Quantization
SURF	Speeded Up Robust Features
BCE	Binary Cross-Entropy

CNNs	Convolutional Neural Networks
MAMD	Multi-Aspect Metric Discrimination
CAC	Central Difference Convolution
STDN	Spoof Trace Disentanglement Network
MT-FAS	Meta-Teacher FAS
DA	Domain Adaptation
DG	Domain Generalization
MMD	Maximum Mean Discrepancy
PCA	Principal Component Analysis
RNNs	Recurrent Neural Networks
LSTM	Long Short-Term Memory
GANs	Generative Adversarial Networks
DDPMs	Double-Diffusion Probabilistic Models
ROI	Region of Interest
SE	Squeeze and Excitation
GAP	Global Average Pooling
EER	Equal Error Rate
HTER	Half Total Error Rate
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
CS	CASIA-SURF
HOOF	Histogram of Oriented Optical Flow

TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
BiD	Bi-directional temporal difference
O	OULU-NPU
M	MSU-MFSD
C	CASIA-MFSD
RA	Replay-Attack
FRR	False Rejection Rate
FAR	False Acceptance Rate
APCER	Attack Presentation Classification Error Rate
BPCER	Bona Fide Presentation Classification Error Rate
ACER	Average Classification Error Rate
SOTA	State-of-the-Art methods

CHAPTER 1

INTRODUCTION

In recent years, biometric authentication has emerged as a superior alternative to conventional password-based security, signifying a transformative shift from outdated practices. With the advent of computer vision and biometric technology [4], individuals can now be reliably identified without relying on credentials or physical artifacts. This seamless integration into daily life has revolutionized critical domains such as airport security, mobile phone authentication, and access control systems [5]. Biometric systems automatically recognize individuals based on their biological and/or behavioural characteristics [6], reducing dependency on cumbersome authentication methods like passwords and tokens. By eliminating the risks associated with forgotten passwords or misplaced cards, these systems offer a user-friendly, efficient, and secure means of authentication.

Among various biometric traits, facial recognition (FR) has become particularly prevalent, as it offers a secure, intuitive, and contactless means of verifying identity. However, the increasing deployment of FR systems has also introduced notable security concerns, especially their susceptibility to presentation or spoofing attacks [7]. Spoofing attacks involve the use of fake artifacts [8] such as printed photographs, 3D masks, or replay videos, with alarming success rates reported at approximately 70% [9]. Ensuring the reliability and security of FR systems [10] is crucial for their application across industries such as forensics, banking security, healthcare, and smart device access. Consequently, the development of face anti-spoofing (FAS) systems to detect and mitigate these attacks has become a critical area of research.

A scientometric review covering the period from January 2015 to April 2024 highlights the continuous research interest and persistent efforts devoted to overcoming challenges within the FAS domain [11]. Figure 1.1 illustrates this trend by showing the steady rise in publications on FAS since 2017, reflecting growing interest and advancements in this field. As biometric technology continues to evolve, addressing security vulnerabilities and enhancing the reliability of FR systems is vital

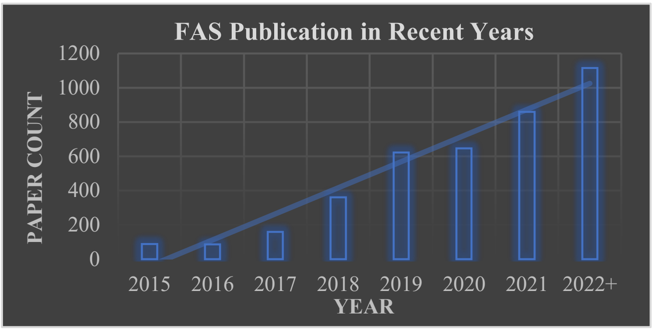


Figure 1.1 Publications over recent years in the field of Face Anti-Spoofing (FAS) obtained through Google Trends.

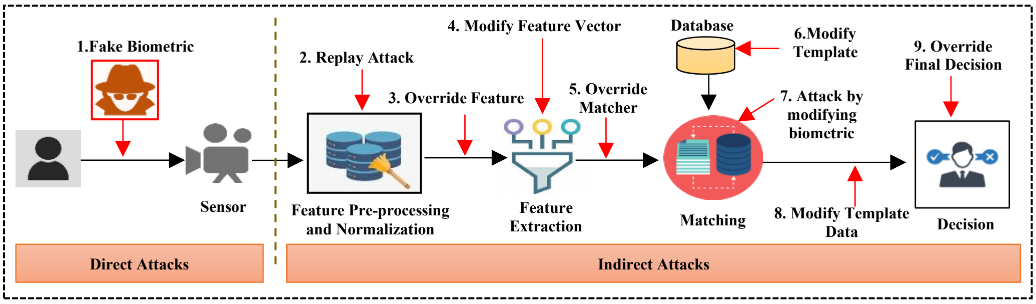


Figure 1.2 Direct and indirect attacks on FR security system [1]

to maximize their effectives while safeguarding against potential threats.

This chapter introduces the foundational concepts of face spoofing attacks and explores the challenges associated with face anti-spoofing techniques. In the final section, the major research contributions of this thesis are discussed, followed by an overview of its organization.

1.1 Face Spoofing and Anti-Spoofing: An Overview

The widespread adoption of face recognition (FR) technology is a double-edged sword—offering convenience while introducing new security vulnerabilities. Spoofing techniques enable intruders to bypass FR-based authentication systems, typically classified into two categories: direct and indirect attacks [1], as depicted in Figure 1.2. Direct attacks are predominantly carried out at the sensor input level and are more prevalent owing to their straightforward execution. They typically involve physical artifacts, such as masks or printed photos, to fool the system. The increasing availability of high-resolution printers and 3D fabrication tools has further exacerbated the realism and frequency of such attacks. In contrast, indirect attacks focus on deeper components of the FR system—like feature extraction, matching algorithms, or databases—and require a more in-depth understanding of the system. However, such

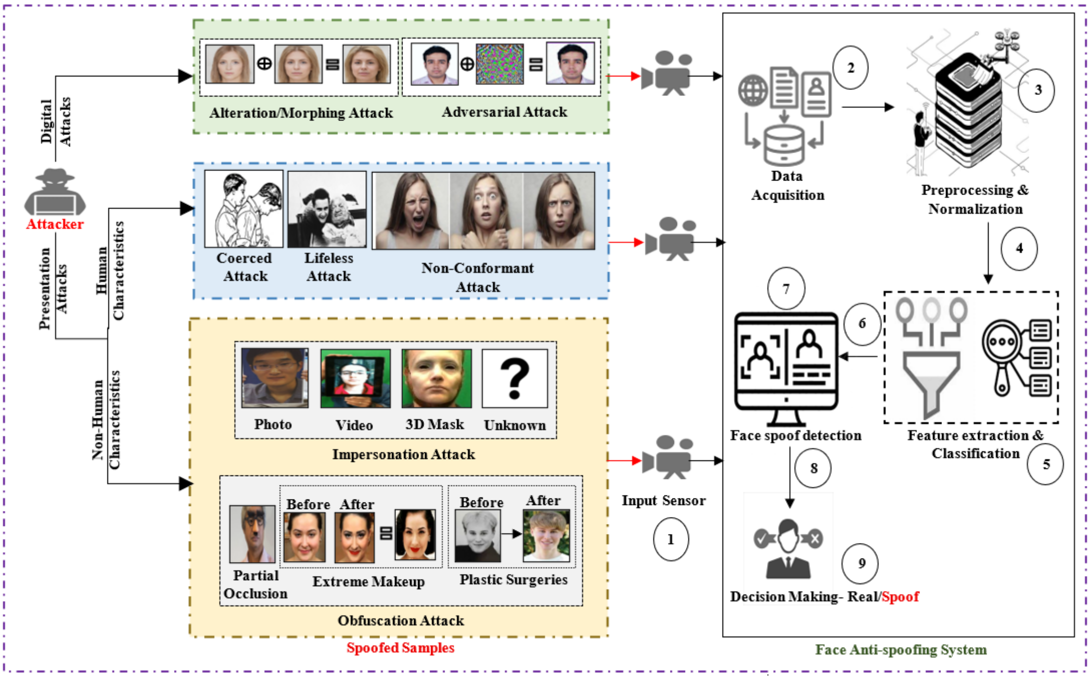


Figure 1.3 Types of Face Attacks

attacks are less frequent, as real-world attackers usually have access only to the input stage.

Direct attacks can be sub-categorized into **digital manipulation** [12] —where spoofing is performed via synthetic modifications in the digital domain—and **physical presentation attacks (PAs)**, wherein tangible spoofing instruments [13] are presented to the sensor (Figure 1.3). In real-world deployment scenarios, physical PAs represent a more immediate and practical threat due to the typical access constraints of adversaries.

Physical PAs can be classified based on the nature of the spoofing instrument into two overarching categories:

1.1.1 Human Characteristics

These attacks utilize a real human subject or their biological traits to deceive the system. They are categorized as follows:

- ***Lifeless Attacks**:* Utilization of a deceased individual's facial biometrics.
- ***Coerced Attacks***: Presentation of an unconscious or unwilling subject’s face.
- ***Conformant Attacks***: Voluntary cooperation of the subject in the spoofing process.

- ***Non-Conformant Attacks****:* Use of deliberate facial movements or expressions to manipulate system behavior.

1.1.2 Non-Human Characteristics

These involve the use of synthetic or manipulated artifacts and can be divided into **impersonation** and **obfuscation** attacks.

1.1.2.1 Impersonation Attacks: With the rise of social media, attackers have easier access to personal photos and videos, enabling them to impersonate legitimate users using real or fake biometric artifacts. The subtypes include:

- 2D Attacks: Include printed, cut, or wrapped photos, designed to mimic facial features and expressions.
  - *Printed photos*: Flat, static images presented to the sensor.
  - *Cut photos*: Altered images with openings (e.g., for the eyes) to simulate minimal motion.
  - *Wrapped photos*: Photos bent or shaped to give a pseudo-3D appearance.
  - *Video replay attacks:* High-resolution recordings [8] displayed to emulate live facial dynamics.
- 3D attacks: Utilize masks fabricated from facial scans or photographs [14] [15]. These masks often replicate fine surface textures and geometric depth, making them especially difficult to detect. The proliferation of low-cost 3D scanning and printing technologies continues to increase the prevalence and sophistication of these attacks.
- Unseen/ Unknown attacks**:** Include novel or evolving spoofing strategies such as prosthetic makeup, deepfakes, adversarial perturbations, and other digitally altered biometric content. These pose a unique challenge due to their variability and lack of training data.

1.1.2.2 Obfuscation Attacks: Aim to hide or disguise identity using occlusion (e.g., scarves, sunglasses), facial modifications (e.g., makeup, plastic surgery), or the use of another individual's biometric trait [16]. Although conceptually distinct from impersonation, obfuscation attacks may employ similar presentation attack instruments. Given their practical relevance and direct impact on system integrity, this thesis emphasizes impersonation-based physical presentation attacks in the subsequent literature review.

To mitigate the growing spectrum of spoofing strategies, numerous Face Anti-Spoofing (FAS) methods have been proposed. These approaches can be broadly categorized into:

1.1.3 Hardware-based Methods

These methods incorporate specialized sensors or acquisition devices to capture discriminative biometric cues for liveness verification. They often necessitate user interaction and are generally intrusive. Common subtypes include:

- ***Sensor-based features:*** Leverage inherent characteristics of the capture device (e.g., depth sensors, IR cameras) to differentiate live from spoofed faces.
- ***Blink Detection:*** Monitors involuntary eye movements to confirm liveness [17].
- ***Challenge Response***: Prompt users to perform specific actions (e.g., head rotation, eye closure) to verify responsiveness [18].
- ***Distinct, intrinsic, and detectable properties of a living body***: These include detectable physical, spectral, electrical, or visual traits like elasticity, density, capacitance, resistance, opacity, or color.
- ***Involuntary signals of a living body:*** This involves detecting physiological signals like blood pressure or electric heart activity.
- ***Multibiometric Methods:*** Multibiometric systems enhance security by combining multiple biometric traits to increase spoof resistance. However, such systems are not inherently spoof-proof [19], as compromising one modality may still yield access. Therefore, recent research has explored modality-specific FAS enhancements for multi-biometric systems.

1.1.4 Software-based Methods

These methods leverage algorithmic strategies to detect and mitigate face PA (see Figure 1.3). As depicted in Figure 1.4, software-based approaches have evolved significantly—from early handcrafted techniques to the sophisticated capabilities enabled by deep learning.

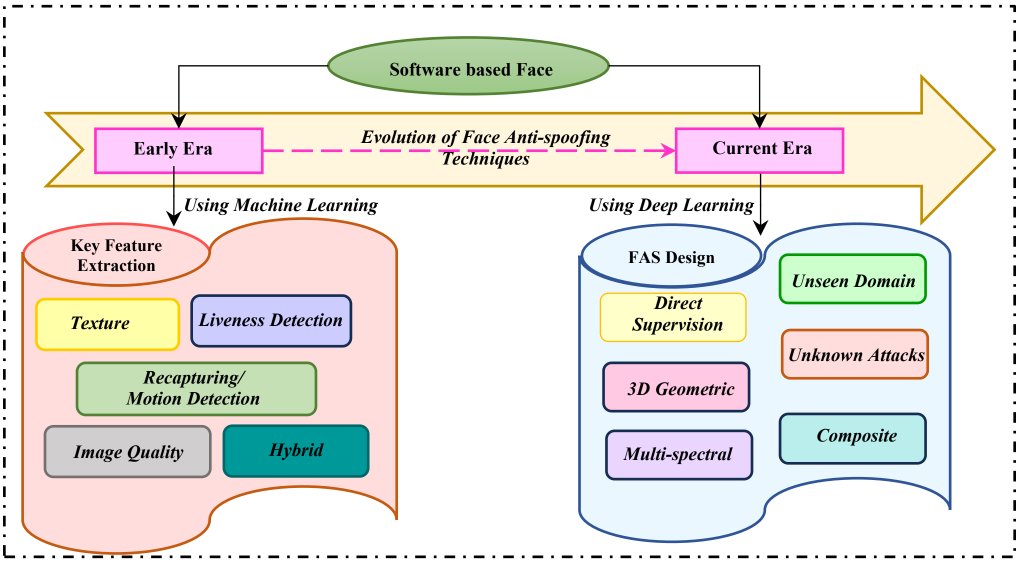


Figure 1.4 Evolution of Software based Face Anti-spoofing methods

Initial methods primarily relied on manually engineered features, such as texture descriptors (e.g., Local Binary Patterns [LBP], Histogram of Oriented Gradients [HOG]), color variance, and motion cues. While these techniques achieved moderate success against basic spoofing attempts, they exhibited limited robustness when confronted with complex or high-quality attacks, largely due to their constrained generalization capabilities.

The advent of deep learning has fundamentally transformed software-based FAS, enabling automated and highly effective detection mechanisms. Instead of relying on manually engineered features, deep learning models operate as adaptive feature extractors trained on large datasets of genuine and spoofed faces. These architectures learn complex, non-linear representations that capture subtle cues such as fine-grained textures, depth inconsistencies, and micro-motion patterns. This paradigm shift has marked a new phase in software-driven FAS, offering superior adaptability, robustness, and resistance to emerging attack types. As a result, modern deep learning solutions provide more reliable and scalable spoof detection, serving as a secure and cost-efficient alternative to traditional hardware-dependent approaches. Figure 1.5 presents a comparison of common framework structures adopted in current FAS systems. Software-based methods are typically categorized into two broad groups:

- ***Static Methods***: Extract features from a single image.

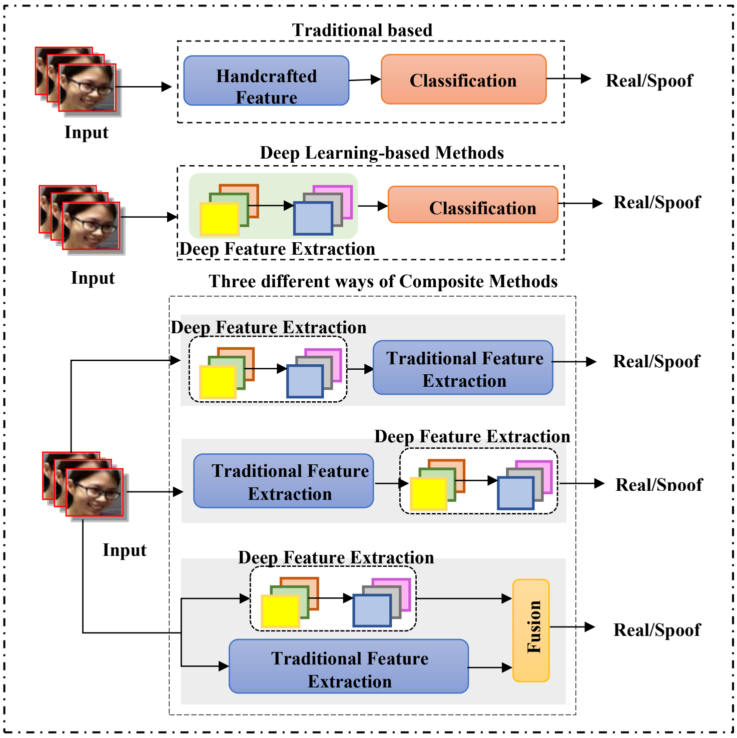


Figure 1.5 Overview of Current Face Spoofing Countermeasures

- ***Dynamic Methods*****:** Gather spatio-temporal features from image or video sequences.

This categorization underscores the diverse approaches used in software-based FAS, each offering unique advantages and capabilities. In this thesis, the literature survey primarily focuses on software-based methods due to their scalability, low deployment cost, and non-intrusive nature. These characteristics make them especially well-suited for real-world applications, including those deployed on mobile and embedded platforms.

1.2 Challenges in the Field of Face Anti-spoofing

Despite significant advancements in FAS, several critical challenges continue to impede the development of robust, generalizable, and deployment-ready solutions. These challenges are as follow:

- Environmental inconsistencies pose a significant challenge in the field of FAS, as they introduce substantial variability in the visual data captured during authentication. Variations in illumination, camera resolution, background complexity, sensor type, pose, and user-to-camera distance can distort or obscure the discriminative cues needed for effective spoof detection. For instance, dynamic lighting conditions alter surface reflectance and shadow patterns, making it difficult to extract consistent texture or depth information. Variations in pose and viewpoint can mask spoof artifacts—such as the edges of printed photos or contours of 3D masks—potentially rendering them undetectable. Background clutter and ambient introduce irrelevant features that divert model attention from the facial region. Additionally, disparities in sensor quality and image resolution further exacerbate the problem by producing inconsistent visual representations, thus complicating cross-device generalization. Varying user-to-camera distances also influence feature granularity, making subtle spoof cues less discernible in low-resolution or distant captures. Collectively, these variations contribute to domain shifts between training and testing environments, ultimately degrading model performance. Addressing these inconsistencies is essential for enhancing the robustness and adaptability of FAS systems in real-world settings.

- Another major challenge in the FAS is the lack of comprehensive and diverse datasets that reflect the full spectrum of spoofing techniques. As presentation attacks evolve—from basic 2D prints to complex 3D masks and multi-modal spoofs—it becomes increasingly difficult to construct datasets that capture both existing and emerging threats. Current datasets often lack diversity, especially in representing 3D and multi-spectral attacks, which are essential for training generalizable models. Furthermore, anticipating future attack methods driven by technological advances adds to the complexity. There is a pressing need for publicly available datasets that mirror real-world variability and include novel attack modalities acquired with modern sensors.

- The development of scalable and efficient multi-modal solutions remains a key challenge in FAS, particularly for real-world deployment. Real-world applications often involve varying combinations of sensor modalities, making it impractical and resource-intensive to train separate models for each possible configuration. Additionally, pseudo-modalities generated via cross-modality translation lack the fidelity and consistency of actual sensor data, further limiting their effectiveness in practical scenarios.

- Ensuring reliable performance in unconstrained, real-world scenarios remains a persistent challenge in FAS. Many existing methods—spanning both traditional handcrafted and deep learning-based approaches—struggle to generalize beyond the conditions observed during training. This often results in overfitting and a corresponding decline in performance when models are exposed to new environments, devices, or users. Such limitations highlight the need for learning mechanisms that can capture intrinsic and discriminative facial features while being invariant to external distortions. Developing robust generalization capabilities is thus critical for achieving dependable performance across diverse operational contexts.
- FAS methods generally perform well against known attack types but often struggle with unseen or unknown attacks due to inherent biases and limited generalization. Addressing this challenge requires adaptive models capable of effectively handling both familiar and novel attack modalities. Traditional FAS approaches lack dedicated mechanisms for identifying unknown threats, while deep learning-based methods offer potential but rely heavily on large and diverse training datasets. Advancing this area remains essential to strengthen FAS systems' resilience and improve security in face authentication applications.

1.3 Research Motivation

The primary motivation behind this study on face spoofing and anti-spoofing stems from the increasing deployment of face authentication systems in real-world applications. According to the International Biometric Group (IBG), face recognition is the second most widely adopted biometric modality in the market [20]. It is extensively utilized in official identification systems, such as national ID cards [21] and ICAO-compliant biometric passports [4].

However, the widespread use of facial biometrics has also heightened the risk of security breaches through presentation attacks, where adversaries exploit spoofing artifacts such as printed photos, video replays, or 3D masks [8].

Several high-profile real-world incidents have demonstrated the critical vulnerabilities of commercial facial authentication technologies. In 2021, CyberArk Labs successfully bypassed Windows Hello—commonly used in professional settings—using a simple USB camera and live images. Microsoft later acknowledged that the attack required no special privileges. Similarly, in 2019, researchers compromised Apple's Face ID using a pair of glasses with tape on the lenses, granting access even when the user was unconscious. Despite Apple's claims of spoofing protection through 3D facial scans, a 2017 demonstration by the Vietnamese firm Bkav revealed that an iPhone X could be tricked using a 3D-printed mask derived from 2D photographs. Additionally, a study conducted by the Consumers' Association Club in the Netherlands found that 40% of tested Android smartphones with face unlock features were susceptible to simple spoofing techniques.

These incidents highlight a clear disconnect between claimed security measures and real-world robustness. The persistent threat of presentation attacks—ranging from simple images to sophisticated 3D masks—highlights the urgent need for advanced, adaptive, and generalizable FAS solutions.

This thesis is therefore driven by the critical need to enhance the security and reliability of face authentication systems, ensuring their safe deployment across diverse real-world environments and attack scenarios.

1.4 Problem Formulation

Face recognition systems have become an integral part of modern authentication and surveillance technologies. However, their susceptibility to presentation attacks poses a critical challenge to their security and reliability. These attacks exploit the inability of conventional face recognition models to effectively differentiate between genuine and spoofed facial inputs, particularly under varying environmental conditions or unseen attack types. This limitation can lead to unauthorized access, highlighting the urgent need for reliable FAS solutions.

This thesis aims to develop and evaluate a deep learning-based framework for the automatic detection of spoofing attacks in facial authentication systems. The

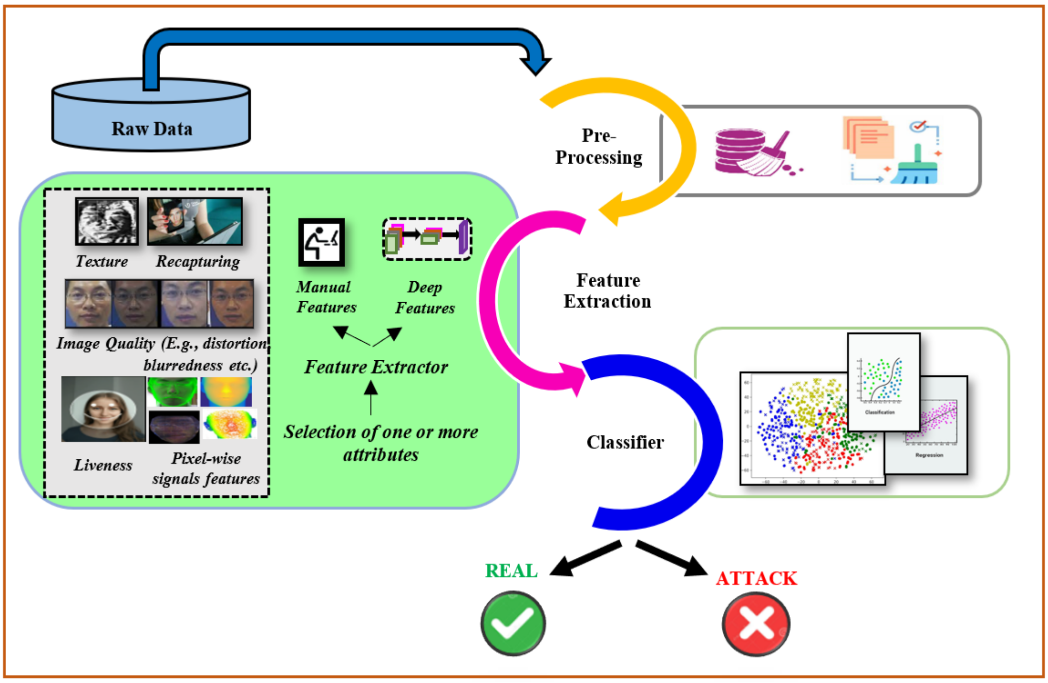


Figure 1.6 Representative diagram to depict the problem formulation approach utilized in this thesis.

proposed system addresses key challenges in presentation attack detection, including generalization across diverse spoofing mediums, lighting conditions, and acquisition devices. The research begins with the collection and preprocessing of face data—comprising both live and spoof samples—validated through benchmark datasets. This is followed by the development of an AI-based system for live/spoof classification and performance evaluation using standard protocols and cross-dataset testing.

Formally, the problem can be modelled as a binary classification task. Let X be the input space containing facial images or video sequences, and $Y = \{0,1\}$ represent the label space, where 1 denotes a live sample and 0 denotes a spoofed sample. The objective is to learn a function $f: X \rightarrow Y$ such that:

$$f(x) = \begin{cases} 1, & \quad\quad live\ samples \\ 0, & \quad\quad fake\ samples \end{cases} \tag{1.1}$$

A representative diagram illustrating this problem formulation is provided in Figure 1.6. The subsequent chapters elaborate on each stage of the system development, from data processing and model design to evaluation and result analysis.

1.5 Research Objectives

The challenges associated with reliable FAS in real-world scenarios highlight the need for efficient and robust algorithms capable of addressing practical limitations. To this end, four key problem statements are formulated to tackle the major challenges outlined above and to guide the development of a resilient and generalizable anti-spoofing system. These are as follows:

- ✓ To propose a deep FAS architecture using RGB images.
- ✓ To implement an efficient multi-modal face anti-spoofing architecture using RGB, depth and IR images.
- ✓ To design spatio-temporal features-based face anti-spoofing architecture for videos.
- ✓ To detect Face Presentation attacks using Pixel wise supervision.

1.6 Research Contribution

The principal objective of this thesis is to address the challenges associated with robust FAS, including variations in illumination, diverse presentation attacks, and the need for effective feature representation across different datasets. To improve generalization and classification performance, this research proposes a transition from traditional single-modal approaches to a multi-modal, generative framework that harnesses complementary information across modalities for more reliable spoof detection. To realize these goals, the following frameworks are proposed:

- A two stream multi-level face anti-spoofing framework has been developed, combining a multi-level ELBP texture feature extraction branch with a modified Xception-based deep feature extraction module. The design aims to capture comprehensive and discriminative face representations while mitigating overfitting by optimizing trainable parameters. The modified Xception network incorporates the squeeze-and-excitation mechanism to enhance multi-level deep features without increasing complexity. Meanwhile, the multi-level ELBP branch optimizes level selection for effective texture feature extraction, balancing model performance and computational efficiency.

- A simple yet efficient multi-modal feature fusion framework, MF^{2}ShrT, has been developed. It integrates overlapping patches and Vision Transformers (ViTs) to enhance local contextual feature extraction. The framework incorporates a parameter-sharing mechanism within the base ViT, termed SharLViT, which strengthens feature representation while significantly reducing the number of parameters and computational complexity. Additionally, a novel T-Encoder-based Hybrid Feature Block is introduced to capture inter-modal correlations and dependencies, ensuring richer feature representations. To further enhance fusion efficiency, an Adaptive Weighted Fusion, and Classification Block (AWFCB) dynamically fuses RGB, Hybrid, and RID branches by learning modality-specific weights, prioritizing salient multimodal features while minimizing redundant information.
- A bidirectional temporal difference-based framework, Bi-STAM, has been developed to capture temporal motion information from both forward and backward perspectives, extracting valuable motion cues to enhance FAS performance. The framework incorporates a Spatio-Temporal Adaptive Modeling (STAM) block, consisting of a Temporal Adaptive Block (TAB) and a Spatial Adaptive Block (SAB). TAB employs an adaptive fusion strategy to balance static semantic and dynamic motion information while effectively learning motion trends. SAB models' appearance features and generates a dense attention map to emphasize semantically relevant attributes critical for FAS. To refine feature integration, the Feature Aggregator Block (FAB) uses an adaptive fusion mechanism to dynamically combine SAB and TAB outputs, optimizing weights to highlight salient features and minimize redundancy, ensuring a robust and efficient FAS framework.
- A novel PolarSentinelGAN FAS framework has been developed, utilizing a dual Polarized Attention-guided Generative Adversarial approach. This approach leverages the complementary information of RGB and MSRCP representations to generate high-quality depth maps. The framework

features a Dual Polarized Self-Attention Guided Module (DPAttn), which adaptively prioritizes RGB and MSRCP features to guide a U-Net-based generative adversarial network, ensuring the generation of distinct depth maps for real and spoof faces. To integrate DPAttn-guided features seamlessly into the generator, Feed Forward Blocks (FFBs) are introduced at each level of the encoder and decoder, facilitating a cohesive fusion of dual features and enhancing overall anti-spoofing performance.

1.7 Significance of the Study

The increasing reliance on facial recognition systems in critical applications such as smartphone authentication, border security, banking, and surveillance has intensified the need for robust FAS techniques. Spoofing attacks—such as printed photos, replayed videos, and 3D masks—can easily deceive facial recognition systems if not properly defended against. This poses serious security risks, particularly in systems where facial biometrics serve as the primary or sole means of authentication.

This study addresses the urgent need for advanced and intelligent countermeasures against such attacks by exploring and developing deep learning frameworks that can generalize well across varied environmental conditions, spoofing mediums, and datasets. Furthermore, the study aims to overcome key limitations in the current state of research, such as poor cross-dataset generalization and high sensitivity to variations in illumination, pose, and camera quality, by proposing innovative and adaptable deep learning solutions. Consequently, the research holds significant theoretical value and practical relevance, contributing to the advancement of secure, AI-driven facial recognition technologies in real-world applications.

1.8 Outlines of the Thesis

The thesis entitled, **‘Deep Learning Frameworks for Face Anti-spoofing’** comprises seven chapters followed by conclusions, future scope, social impact, and a bibliography. The thesis is organized as following:

Chapter 1: Introduction

This chapter provides a comprehensive overview of face-spoofing attacks, examining both traditional and emerging face anti-Spoofing (FAS) techniques from leading research. It establishes the foundational knowledge necessary for subsequent chapters, highlighting key challenges and underscoring the need for advanced FAS solutions.

Chapter 2: Literature Survey

In this chapter, a systematic review state-of-the-art face anti-spoofing (FAS) methodology is presented. It analyzes backbone architectures, feature extraction techniques, and dataset dependencies while examining their robustness against diverse attack types. Advancements in domain adaptation, multimodal fusion, and attention mechanisms are explored. The chapter concludes by identifying research gaps and defining the core objectives of this thesis.

Chapter 3: Two-stream RGB-based FAS framework

This chapter presents an innovative RGB-based face anti-spoofing framework that utilizes a two-stream multi-level architecture. It combines multi-level Elliptical Local Binary Patterns (ELBP) texture features with a modified Xception network enhanced by Squeeze-and-Excitation (SE) modules. This integration optimizes both texture and deep feature extraction, leading to robust spoof detection capabilities. The chapter delves into key aspects such as feature selection strategies, network optimization, and performance trade-offs, laying a foundational groundwork for deep learning-based face anti-spoofing solutions.

Chapter 4: Multimodal Vision Transformer for Robust Face Anti-Spoofing

This chapter introduces the MF^{2}ShrT multimodal face anti-spoofing framework designed for enhanced resilience against sophisticated presentation attacks. The framework employs a Vision Transformer (ViT)-based architecture with overlap patches and parameter sharing technique to improve computational efficiency. It also incorporates a T-encoder-based hybrid feature block that effectively captures cross-modal dependencies. By optimizing feature fusion across different modalities, the framework achieves highly discriminative and robust spoof detection capabilities.

Chapter 5: Spatio-Temporal Adaptive Modeling for Face Anti-Spoofing

This chapter introduces the Bidirectional Spatio-Temporal Adaptive Modeling (Bi-STAM) framework for face anti-spoofing, which enhances robustness against dynamic attacks by leveraging motion cues from both forward and backward temporal perspectives. The proposed Spatio-Temporal Adaptive Modeling (STAM) integrates two key components: a Temporal Adaptive Block (TAB) to learn dynamic motion trends and a Spatial Adaptive Block (SAB) to refine appearance-based representations. Additionally, a Feature Aggregator Block (FAB) employs an adaptive fusion mechanism to balance static semantics with motion dynamics, thereby improving the discriminative performance of video-based anti-spoofing systems.

Chapter 6: Generative Learning-Based Pixel-Wise Face Anti-Spoofing Framework

This chapter presents a GAN-driven face anti-spoofing framework that utilizes pixel-wise supervision to enhance the discrimination between real and spoofed faces based on depth cues. The framework introduces PolarSentinelGAN, which incorporates a Dual Polarized Self-Attention Guided Module (DPAttn) to optimize the embedding of RGB and MSRCP features. Additionally, the chapter explores the use of a hierarchical Feed Forward Block (FFB) within an encoder-decoder architecture, facilitating structured refinement of features and ensuring robust generalization across diverse attack scenarios.

Chapter 7: Conclusion, Future Scope, and Social Impact

This chapter synthesizes the thesis's key findings and contributions, evaluating the effectiveness of proposed methodologies in advancing face anti-spoofing. It analyzes real-world deployment impacts, discusses potential enhancements, and outlines future research directions. The chapter also explores broader societal implications, including ethical considerations and the role of AI in enhancing biometric security.

CHAPTER 2

LITERATURE REVIEW

This chapter presents a comprehensive review of the state-of-the-art methodologies in face anti-spoofing (FAS), with a particular focus on the evolution of backbone architectures and the datasets employed in the field. To trace the evolution of various FAS techniques over time, the existing literature is broadly categorized into two primary domains: fundamental key features for FAS and FAS design approaches.

2.1 Fundamental Key Features for FAS

Earlier, researchers explored diverse feature types to address the inherent vulnerabilities of face authentication systems. These features have been systematically classified into five categories based on their extraction principles: texture-based, image quality-based, recapturing-based, liveness detection-based, and hybrid approaches. Despite their effectiveness, the development of these detection algorithms often necessitates substantial prior knowledge and expertise from researchers.

2.1.1 Texture methods

These methods differentiate between spoofs and genuine faces based on the differences in their texture or illumination patterns. A range of operators, such as LBP [22], HOG [23] and LPQ [24]are used to extract texture information, fed as input to supervised classifiers like SVM or LDA for binary classification [25]. Early studies focused on grayscale texture analysis using LBP to extract texture details. However, this method was only effective for high-resolution, texture-clear deceptive face images and failed for low-resolution deceived face images. LPQ-based color texture analysis [24] was initially used but had issues with sensitivity to lighting changes and high-quality data. To further enhance the discrimination, advanced techniques such as SURF [26], frequency domain energy analysis [27], and edge information were utilized by the researchers. Some methods also incorporated temporal features from the sequence of frames or used edge information for texture representation. Despite their promising results, these methods are sensitive to illumination conditions, electronic displays, and occlusions, and require substantial pre-processing of data, making them less efficient than other approaches.

2.1.2 Recapturing-based methods

These methods detect whether the biometric sample is a replay of recorded data. This method examines the scene and environment for abnormal movements caused by a hand holding a 2D electronic device or paper, which differ from actual face movements against a static background [28]. These methods may use an optical flow or motion analysis to track relative motion between facial parts, visual rhythm analysis, motion magnification, or Haralick features. Contextual analysis [23] of the scene or 3D reconstruction may also be employed. However, these methods are not advantageous against attacks involving an artifact such as a mask due to absence of a replay, there is nothing to detect. A more effective and versatile method is texture-based approaches that use traits to distinguish genuine biometrics from replay data or artifacts presented to the sensor.

2.1.3 Image Quality Analysis

Deceiving the presentation of a human face, whether on picture paper, printing paper, silica gel, or electronic devices, requires the utilization of certain media. However, the material properties of these artifacts do not align with those of an actual individual's face, leading to variances in reflection quality. Materials like picture paper and mobile phone display screens exhibit specular reflections but lack the characteristics of living faces. This method observes image distortion [8] commonly found in spoofed face images like print attacks where the face may skew or deformed due to bending. Frequency analysis [29] is another way that considers how previously captured images or videos may affect the frequency information of the sample when displayed in front of a sensor. This approach may fail with high-quality spoof images or videos. Some experts [7] [30] explore quality metrics-based methods with the assumption that the display device or paper distorts various picture quality measurements. Apart from this, many experts used color analysis [26], where the distribution of color in images captured from bonafide faces may differ from those taken from display devices or printed papers. Methods for face PAD adopt solutions in different input domains, such as HSV and YCbCr color space, Fourier spectrum [31], or color adaptive quantized patterns [26]. Overall, it is a powerful tool to detect distortion in spoofed face images.

2.1.4 Liveness Detection

Many authors have observed that most spoof attacks happen using still images that lack any basic motion. Thus, this approach involves analysing signs of life, such as blinking eyes [32], changes in facial expressions, and movements of the head [33]. As the living face is dynamic, this method exploits temporal features. This approach further divides into active/intrusive [34] and passive/non-intrusive [35] [36]. The former requires user interaction, while the latter does not require user cooperation, such as spontaneous eye blinking or unconscious facial reactions. However, these approaches can be vulnerable to spoof attacks, such as replay videos or the addition of liveness properties to a still image. The entire verification process usually involves long-term interaction, making it unfavourable for practical implementation. In contrast, RPPG [37] is an extensively used methodology that can detect photo attacks, low-quality 3D masks, and lower-quality replay attacks but is less reliable against high resolution replay PAs that display dynamic variations of the live person.

2.1.5 Hybrid-based methods

Choosing a single technique to counter the growing diversity and realism of spoofing attacks has become challenging. As a solution, many experts are now hybridizing multiple-face anti-spoofing methods to improve generalization ability. One such method is the fusion of image quality and motion cues [38], accomplished through feature-level fusion employing low-level feature descriptors with partial least squares regression [39].

2.1.6 Discussion, challenges, and limitations

The key features identified for FAS methods were subsequently categorized under traditional-based methods, involving manual design and selection of features to detect facial spoofing attempts. While these countermeasures have demonstrated efficacy in certain scenarios, discussions around traditional-based methods often revolve around interpretability and the use

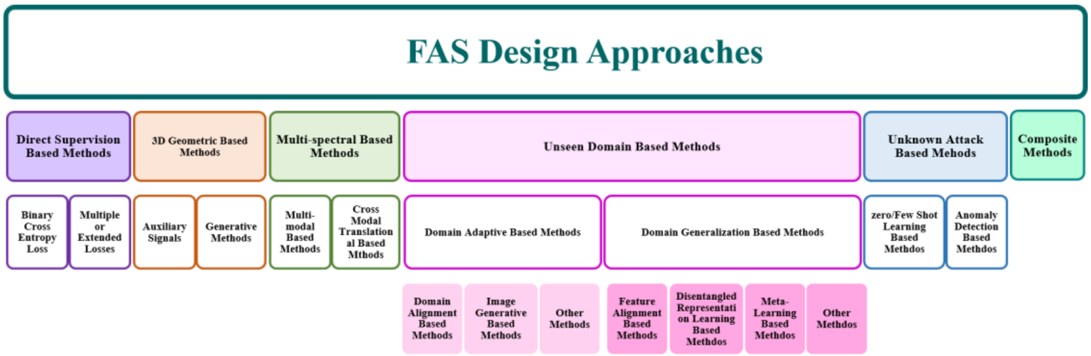

FAS Design Approaches
Direct Supervision Based Methods
3D Geometric Based Methods
Multi-spectral Based Methods
Unseen Domain Based Methods
Unknown Attack Based Mehods
Composite Methods
Binary Cross Entropy Loss
Multiple or Extended Losses
Auxiliary Signals
Generative Methods
Multi-modal Based Methods
Cross Modal Translational Based Mthods
Domain Adaptive Based Methods
Domain Generalization Based Methods
zero/Few Shot Learning Based Methdos
Anomaly Detection Based Methdos
Domain Alignment Based Methods
Image Generative Based Methods
Other Methods
Feature Alignment Based Methods
Disentangled Representation Learning Based Methdos
Meta-Learning Based Methdos
Other Methdos

Figure 2.1 Classification of FAS Design Approaches

of domain-specific knowledge to enhance anti- spoofing capabilities. Despite their utility, traditional-based methods come with certain challenges. It includes the need for domain expertise in identifying relevant features and limitations arising from fixed parameters hindering adaptation to various spoofing attacks. These methods may struggle to capture the complexity of facial variations in diverse datasets and real-world scenarios, and with their manual design being time-consuming and less adaptable to evolving spoofing methods. Limitations also encompass susceptibility to adversarial attacks and difficulties in addressing rapidly evolving spoofing techniques. Despite these limitations, traditional-based methods remain widely used, offering computational efficiency, easy implementation, and interpretation. They serve as a good representation of facial features and can be integrated with other countermeasures, such as deep learning techniques, to create robust systems. Notably, the research in this area is ongoing with continuous development and evaluation of new key features and techniques. Staying abreast of the latest developments is essential for evaluating their performance in specific applications.

2.2 FAS Design Approaches

In this section, we will review various deep learning-based approaches employed to design FAS systems, as shown in Figure 2.1. This serves as the basis for organizing and presenting these insights.

2.2.1 Direct supervision-based methods

The field of FAS has evolved significantly from manual feature extraction

[24] [40] [41] to deep learning techniques [42] [43] [44] [45]. Traditional methods relied on manual extraction of specific facial features, limiting their adaptability to emerging spoofing techniques. However, the emergence of deep learning paradigms has marked a revolutionary shift for FAS. Unlike their predecessors, deep learning models eliminate the necessity for manual feature identification. Instead, they utilize intricate architectures and regularization strategies to autonomously discern complex patterns directly from the data. This paradigm shift has propelled deep learning to the forefront of FAS research, offering more adaptable and effective solutions to combat spoofing attacks. Initially, these deep learning models approached FAS as a binary classification problem (e.g., '0' for real and '1' for spoof faces), employing binary classifiers or extended losses for supervision.

Binary Cross-Entropy Loss: Binary Cross-Entropy (BCE) loss used to be a key component in many FAS systems, lauded for its simplicity and efficacy. However, its utility was constrained by shallow convolutional neural networks (CNNs), which struggled with diverse datasets and tended to overfit [46]. To surmount these challenges, researchers turned to transfer learning, employing pre-trained models like VGG16, VGG11, and ResNet50 to bolster FAS capabilities [47] [48] [49] [50]. While effective, these approaches were time-consuming due to their reliance on transfer learning. Moreover, despite excelling at capturing high-level features, these models often overlooked critical low-level details crucial for detecting spoofing patterns [51]. Recognizing the significance of learning deep local features from each facial region, Souza et al. [52] proposed LSCNN, demonstrating how learning different local spoofing cues can enhance performance. Chen et al. [53] addressed the impact of varying lighting conditions on FAS performance by dividing images into RGB and Multi-Scale Retinex (MSR) spaces and developing a two-stream CNN network called TSCNN. Meanwhile, Deb et al. [54] introduced SSR-FCN, aiming for generalizability and interpretability against 13 different spoof types. A new paradigm emerged with ViTranZFAS [55] integrating pre-trained vision transformer models into FAS tasks for the first time. Although effective, ViT's demand for massive pre-training datasets poses a time-consuming challenge.

Beyond static appearance, the exploration of temporal dynamics through

LSTM [49], GRU [56], RNN [57] networks gained momentum, enabling the detection of subtle temporal inconsistencies between real faces and spoof presentations, thus enhancing FAS robustness in real-world scenarios. However, as FAS research progressed, limitations of BCE surfaced. Class imbalance in real-world scenarios, with fewer spoof attempts than real faces, can cause the model to prioritize learning patterns for real faces, potentially overlooking spoof attempts. Additionally, BCE treats all spoof types equally offering no insights into specific spoofing types the model might struggle with.

Multi-class or Extended losses: Researchers, recognizing the limitations of BCE loss, have explored advanced loss functions to address class imbalance and provide detailed feedback on spoofing types, thus fostering the development of more robust FAS systems. These extended loss functions surpass binary classification, incorporating supplementary information to enhance model resilience and versatility. Numerous studies [58] [59] [60] [61] [62] [63] have proposed modifications to enhance feature discrimination in FAS tasks. For example, PatchCNN [64] adopts a multi-objective approach by integrating BCE and Triplet losses to improve feature discrimination. CIFL [63] utilizes Binary Focal loss alongside camera type data to enhance the network's ability to detect camera-invariant spoofing features. PatchNet [65] introduces the Asymmetric AM-Softmax Self-supervised Similarity loss, emphasizing strong regularization in the patch embedding space for fine-grained live/spoof recognition. MTSS [66] exploits CE loss in Vision Transformer (ViT) for Multi-Aspect Metric Discrimination (MAMD), augmenting feature discriminability and robustness. Lastly, Qiao et al. [67] implement a 5-class Cross-Entropy (CE) loss in an efficient fine-grained detection network using a Transformer-style model, integrating data augmentation and convolution self-attention mechanisms.

In general, BCE loss and its variations serve as convenient and effective tools for training deep FAS models. However, their reliance on spatial or temporal global constraints for live/spoof embedding learning can sometimes result in inaccurate patterns, overfitting, and decreased architectural performance. Consequently, researchers have delved into alternative approaches, such as leveraging pixel energy contents, to guide networks in effectively tackling face spoofing detection challenges.

2.2.2 3D Geometric based methods

While loss functions like BCE have played a crucial role in training deep learning models for FAS, they primarily focus on the final classification outcome (real vs. spoof). However, achieving robust spoof detection necessitates the model to learn more nuanced details about the underlying facial patterns. This is where 3D geometric based methods step in. They offer a complementary approach to loss functions by directly guiding the model's learning at the pixel-level. Such supervision can be attained either through auxiliary signals or generative methods.

Auxiliary Signals*:* In the quest for enhanced FAS, auxiliary signals emerged as indispensable allies by providing additional cues crucial for discerning genuine faces from spoofed presentations. Initially, Wang et al. [68] pioneered photo attack detection by constructing 3D data from 2D planar spoofs captured at varied viewpoints. However, the necessity for multiple viewpoints and the lack of a keyframe rendered this approach unviable [69]. Subsequently, attention shifted towards auxiliary signals-based methods, delving into patch-level cues, and offering context-aware supervision at the pixel level. These methods empower networks to autonomously identify pivotal regions for accurate face detection, leveraging diverse features such as depth maps [70] [71] [72] [73] [74], reflection details [75] [76], binary mask labels [77] [78] [79]and 3D cloud maps [69] for spoof classification. While spoofing attempts lack genuine depth, Atoum et al. [73] for spoof classification pioneered the use of pseudo-depth labels derived from estimated distances between the camera and facial points, revealing depth irregularities associated with spoofing. Subsequent research [70] [74] [80] utilized these labels to predict depth maps for live faces and zero maps for potential attacks. Wang et al. [81] incorporate deep spatial gradient and temporal information to assist depth map regression, while Yu et al. [70] propose a novel frame-level FAS method based on Central Difference Convolution (CDC), leading to the development of widely used architectures like DepthNet/CDCN [82] [83] [84]. Peng et al. [74] amalgamated depth-enhanced networks with color streams to achieve resilient features. Conversely, AISL [85] utilized DepthNet for smooth HTER into a differentiable function. TTN [86] explores vision transformers [87] for temporal-based auxiliary signal architecture, breaking new ground. Despite the benefits, challenges such as expensive and imprecise 3D shape synthesis prompt exploration of alternatives like binary mask labels [78] [88] [89] [90] [91] which segment facial regions into binary representations, aiding spatial localization and isolation from surroundings. Liu et al. [78] expanded zero-shot FAS to diverse spoof attacks using binary supervision and pixel-wise mask regression, whereas Yu et al. [92] sought lightweight FAS architectures with pixel-wise binary supervision. Additionally, Li et al. [69] utilized sparse 3D point cloud maps for efficient supervision, and Liu et al. [93] design a CNN-RNN model to leverage the Depth map and Rppg signal as supervision. Similarly, Yu et al. [92] treated FAS as a material recognition problem and combine it with classical human material perception, aiming to extract discriminative and robust features for FAS tasks. Other auxiliary signals like pseudo-reflection maps [94], Fourier spectra [95], and ternary maps [96] have also shown promise in deep FAS. Integrating these diverse supervisions through multi-task learning offers potential advantages. Further research is needed to determine whether these signals consistently capture intrinsic features that differentiate real faces from spoof attempts. While auxiliary signals [97] [98] offer a powerful approach, its performance hinges on the availability of high-resolution training data and the reliability of the generated pseudo-labels.

Generative methods*: G*enerative methods like Generative Adversarial Networks (GANs) [99] [100] and diffusion models [101] are emerging as powerful tools against spoofing attempts. By generating synthetic spoofing samples, they address challenges in dataset diversity, enriching training data with varied attack scenarios and facial characteristics. As a result, they enhance the resilience of FAS systems in real-world scenarios. Despite their operation at the pixel level for image generation and analysis, generative-based FAS models [102] [103] do not require explicit pixel-level supervision for spoof detection. As a result, they are often treated as a distinct approach from pixel-level supervision methods in FAS, though some works [13] [104] may categorize them under this category.

- GAN-based FAS models harness the power of adversarial training, offering innovative solutions to the challenges of spoof detection. For instance, Jourabloo et al. [102] pioneered a unique methodology by decomposing spoof faces into live faces and spoof noise patterns, utilizing the latter for classification purposes. Similarly, Stehouwer et al. [105] introduced a GAN-based architecture capable of synthesizing and identifying noise patterns across various medium/sensor combinations, including both known and unknown scenarios. Liu et al. [106] further advanced this field with the Spoof Trace Disentanglement Network (STDN), employing U-Net architecture and Patch GAN to disentangle spoof traces from input faces across multiple scales. Despite these advancements, challenges persist, particularly in addressing trace and feature diversity. Zhang et al. [107] proposed an innovative approach to disentangle liveness and content features from images. This approach enhances classification accuracy by employing a CNN architecture with multiple supervisory signals. Meanwhile, Qin et al. [108] introduced the Meta-Teacher FAS (MT-FAS) method, training a meta-teacher for supervising PA detectors. To overcome limitations in generalization and adaptability to unseen attacks, Wang et al. [109] developed a dual-stage disentangled representation learning method, improving identification precision. Nevertheless, the field still grapples with the challenge of learning from limited data variations, as evidenced by initiatives such as the PhySTD [99] network, aimed at disentangling spoof traces from input faces. Despite these advancements, existing models often struggle to generalize effectively across diverse environmental conditions and spoofing materials. To address this issue, Wang et al. [104] propose employing GANs to transfer input face images from the RGB domain to the depth domain, albeit facing challenges in training and diversity. In response, Zhang et al. [110] proposed a novel FAS scheme utilizing Wasserstein GAN to enhance the depth transfer network. These developments underscore the ongoing evolution of GAN-based FAS methods, highlighting the continued potential for exploration and innovation in this field.

- Diffusion-based methods present an alternative approach, employing diffusion processes to identify spoof attempts by scrutinizing minute variations in image characteristics. For instance, Alassafi et al. [103] introduce a pioneering face presentation attack detection (PAD) solution utilizing interpolation-based image diffusion augmented by transfer learning from a MobileNet convolutional neural network. Furthermore, in an extended application of this technique, Zhang et al. [101] leverage the diffusion model to segregate noise patterns within spoof images and reconstruct the authentic counterparts. Although research on diffusion models in FAS is still nascent, preliminary investigations show promising outcomes.

Research persists in the quest for optimal auxiliary signal generation to differentiate spoofed data from genuine faces. Although GAN-based and diffusion models hold promise, challenges such as data quality and adaptation to evolving spoofing techniques persist. Improving model generalizability to detect unseen/unknown attack types remains a priority for robust FAS systems. Ongoing research indicates that generative models will significantly impact the future security of facial authentication systems.

2.2.3 Unseen Domain Based Methods

Although techniques incorporating supervision, 3D geometric, and modality have demonstrated impressive performance, they often struggle with unfamiliar domain conditions like lighting variations, backgrounds, and facial features. This unreliability stems from domain distribution gaps and overfitting, making them unsuitable for practical applications. Consequently, domain adaptation (DA) and domain generalization (DG) have emerged as main strengths for FAS. These methods enable models to effectively generalize on unseen testing domains by learning from diverse but interconnected training datasets. The only difference between the two techniques is that DA has access to the target domain data while DG does not during training. Although this makes DG more challenging, it is more favourable and complementary in real-world scenarios.

Domain adaptive based methods (DA)**:** This approach aims to transfer knowledge acquired from a source domain with ample labeled samples to a target domain consisting solely of unlabeled data. The primary objective is to minimize the distribution gap between the source and target domains, enabling the learned models to effectively adapt to the target domain. Depending on the specific method employed, existing approaches can be categorized into three sub-groups: domain alignment-based, image generation-based, and other methods.

Domain Alignment-Based methods: Domain adversarial learning and maximum mean discrepancy (MMD) minimization are widely employed domain alignment-based methods in FAS. Initially, Li et al. [111] proposed learning domain-invariant features by minimizing the MMD between source and target domain feature distributions, marking an early approach in this domain. Subsequently, domain adversarial learning gained prominence due to its effectiveness. Wang et al. [112] utilized domain adversarial learning to guide the feature encoder in mapping source and target domains to a shared feature space where domain labels are indistinguishable. Later, [113] integrated disentangled representation learning, metric learning, and domain adversarial learning to separate domain-specific features and align domain-invariant ones within the shared feature space. Eldin et al. [114] combined deep cluster learning with domain adversarial learning to retain the specific characteristics of the target domain while acquiring domain-invariant features for accurate classification. Panwar et al. [115] introduced compound DA for FAS, treating the target domain as a combination of multiple similar domains without explicit domain labels. Recognizing the feasibility of obtaining labeled samples in some face anti-spoofing applications, Jia et al. [116] proposed a unified framework for both unsupervised and semi-supervised learning.

Image Generation-Based methods: Tu et al. [117] employ a technique where images from the target domain are transformed into synthetic images with styles akin to those in the source domain, aiming to reduce the domain gap. Acknowledging the diverse acquisition environments and capturing devices present in public face recognition datasets, initially [118] employs an autoencoder to capture these variations. Subsequently, it generates reconstruction-error images for training FAS models, promoting the model's insensitivity to such variations and enhancing its generalization capability.

Other methods: Neural network pruning and knowledge distillation techniques have also been employed to enhance the generalization performance of FAS methods. Neural network pruning and knowledge distillation techniques have been applied to enhance the generalization performance of face anti-spoofing models as well. Mohammadi et al. [119] discovered that certain filters within a specific convolutional layer exhibit domain-specific characteristics, while others are domain-invariant. Leveraging this insight, they prune domain-specific filters based on computed feature divergence values, thereby enhancing the generalization capacity of pre-trained models. Further, Li et al. [120] introduce a knowledge-distilling framework designed to transfer knowledge acquired by a teacher network, trained with abundant labeled samples from the source domain, to a student network tailored for the target domain, even when only limited training data is available.

While existing DA methods have significantly enhanced the generalization capabilities of FAS models, they suffer from a critical limitation. These approaches assume the target domain utilizes the same PAIs as the source domain. In real-world scenarios, this assumption is often violated, as attackers may employ entirely new spoofing techniques (unknown PAIs). Consequently, these models struggle to generalize effectively to such unseen attacks.

2.2.4 Unknown Attacks Based Methods

The ever-evolving nature of application scenarios and unpredictable emergence of novel PAs pose challenges for data-driven FAS models. These models may struggle with out-of-distribution data, such as real faces captured in new environments or unknown spoofing techniques. Traditional methods rely on extensive labeled data for each new attack, which is expensive and impractical due to rapid spoofing methods evolution. Deep learning approaches often treat FAS as a closed-set problem, assuming a finite number of known attacks, leading to overfitting and poor generalization to unknown attacks. Recent research explores alternative approaches like zero-shot and few-shot learning to address these limitations, enabling FAS models to detect unknown attacks with minimal labeled data. Treating FAS as a one-class classification problem or using anomaly detection methods can enhance spoof detection with limited samples of new attack types, focusing on identifying deviations from the expected distribution of real faces.

Zero/Few-Shot Learning: Detecting new spoofing attacks is vital for FAS system efficacy. Traditionally, fine-tuning the FAS network with numerous new PA samples is common. However, this can be both time-consuming and expensive. To overcome this, researchers have proposed innovative approaches such as zero-shot and few-shot learning. Zero-shot learning allows FAS models to extract features from predefined attacks without new sample training while few-shot learning combines predefined attacks with a small set of new samples for training. For instance, Liu et al. [78] and Qin et al. [121] introduced methods for zero- and few-shot FAS tasks, focusing on grouping spoof samples and training a meta-learner, respectively. However, few-shot meta-learning may cause forgetting of predefined attacks. To address this, Perez-Cabo et al. [122] proposed a few-shot learning approach for continual learning, extending knowledge incrementally. Despite advancements, FAS models face challenges in zero-shot cases, especially with difficult attacks like transparent masks or makeup. Further research is necessary to enhance FAS system robustness against zero-shot scenarios and novel spoofing attacks.

Anomaly Detection*:* This technique assumes genuine samples form a distinct class, while anomalous or outlier samples have significantly different distributions in the sample space. It involves training a one-class classifier to group live samples, flagging any samples outside the genuine cluster as attacks. Anomaly detection leverages the close cluster behavior of live samples, gaining popularity for detecting unseen attacks in PAD. Genuine faces are seen as normal with lower variance, forming a closely clustered group. Attacks display higher variance, leading to anomalies. Several methods [123] [124] have been developed for anomaly detection [125] in FAS, including GMM [126], One-class SVM [123], autoencoder-based outlier detectors with LBP feature extractor [123], pairwise one-class contrastive loss (OCCL) [127], triplet focal loss [124]. One-class classifiers, like Anomaly [123], Anomaly2 [124] and Fatemifar et al. [128] are a popular choice for anomaly detection in FAS. These methods train a model specifically on real face data. These train models on real face data, flagging deviations from this representation as anomalies. Techniques like Metric Learning [129] further enhance this approach by creating robust distance measures in the feature space. End-to-End Learning [130] trains both anomaly classifier and feature representation simultaneously. It trains both the anomaly classifier and the feature representation in a single step. Open-Set Recognition [131] views it as an open-set recognition task, leveraging statistical methods to identify unknown PAs. While offering advantages, anomaly detection may have lower performance in practical open-set scenarios compared to traditional methods. Nonetheless, it remains valuable for FAS systems, offering robustness against new spoofing attempts with minimal training data. As research progresses, anomaly detection can play a crucial role in securing future FAS systems.

2.2.5 Multi-Spectral based methods

Advancements in hardware manufacturing and integration technology have facilitated the growing adoption of cost-effective multi-spectral-based FAS solutions in real-world applications. Relying solely on the visible spectrum for detecting PAs has proven inadequate and challenging. Consequently, auxiliary multi-modal information such as depth and/or infrared (IR) images is now employed to enhance PA identification.

Multi-Modal based methods*:* With the availability of large-scale multi-modal datasets [16] [132] and high-fidelity mask datasets [16], exploring multi-modal FAS tasks is crucial for technological advancement. These methods integrate various data sources beyond traditional RGB images, including depth maps, thermal data, and infrared (IR) data. These modalities offer unique insights into facial characteristics, collectively enhancing spoof detection robustness. Feature fusion [133] [134] combines information from each modality into a unified representation. Therefore, studies have shown the effectiveness of feature-level fusion in multi-modal FAS with architectures like FaceBagNet [135], leverages local image blocks for feature extraction and integrates information through modal feature erasure, effectively utilizing the interplay between modalities. However, existing approaches have limitations. Traditional global fusion strategies [16] [136] [137], while effective, combine information from entire images across modalities. This can overlook crucial localized spoofing artifacts, like subtle imperfections on a mask. Local fusion [135] [138]addresses this by focusing on specific image regions, potentially uncovering these subtle clues. The ideal solution lies in a combined approach [139], capturing both the global context and local variations. Recent work [140] utilizing shared layer ViT for joint local and global feature extraction offer a promising direction for achieving this. However, feature-level fusion [138] [141] frequently necessitates separate branches for extracting modality features, leading to increased computational costs.

Beyond the feature-fusion technique, the stage of fusion also plays a critical role. some works employ decision-level fusion [142] to balance individual modality biases and ensure reliable binary decisions. However, it necessitates separate well-trained models for each modality, making it less efficient. In contrast, input-level fusion [139] [143] [144] offers a more efficient alternative by combining pre-processed data from different modalities before feature extraction. For instance, Wang et al. [145] model exemplifies this approach. It first uses convolutions on each block to extract local features, followed by "cross-sheet mixing" across different patches to capture long-range global correlations. This allows the model to learn the fusion process itself, potentially leading to more optimal feature representations for spoof detection, although interpretability of the learned features might be reduced compared to feature-level fusion.

Cross-modal Translation based methods*:* Some PAD systems employ multiple sensors to capture diverse facial modalities. However, in certain scenarios, only specific modalities are accessible, such as RGB. To overcome this constraint, approaches combine cross-modal translation methods with fusion methods to bridge the semantic gap between different modalities. For example, Liu et al. [141] introduced a novel auxiliary network for cross-modal translation, consisting of MT-Net and MA-Net, which leverages NIR images for effective discrimination between real and spoofed faces. Likewise, Jiang et al. [146] developed a cycle-GAN for image translation, enabling the generation of NIR images from RGB face images and learning fused features. Additionally, Mallat and Dugelay [147] proposed a conversion method from visible to thermal images, facilitating the creation of thermal image samples from RGB face images. However, these methods also face challenges in domain generalization scenarios due to sensitivity to modality, unreliability, and imbalance leading to inadequate resistance against domain shifts.

Multi-modal FAS offers promise in enhancing facial authentication by integrating effective fusion techniques to capture both global and local information, surpassing traditional single-modal approaches. However, ongoing exploration of novel fusion strategies and optimization techniques is crucial to solidify its role in ensuring the resilience and reliability of anti-spoofing systems. Despite its increasing adoption, multi-modal FAS lags RGB-based methods in advancement, highlighting the need for more efficient modal approaches and tailored sensors for FAS applications. Nonetheless, integrating advanced sensors such as light field, SWIR, and polarization may pose challenges in terms of portability and cost for practical deployment.

2.2.6 Composite Methods

Despite notable advancements across various domains, deep neural networks continue to grapple with overfitting in numerous FAS tasks due to the limited availability of small-scale training datasets. Consequently, researchers have delved into composite methods [148] [149] [150] [151] [152] amalgamate traditional feature extraction with deep learning models. The objective is to harness the strengths of both approaches to attain robust and generalizable performance. For instance, one approach [153] involves extracting handcrafted features like LBP or color-based features, then inputting them into a deep network such as a random forest for classification. Another method [38] extracts motion features (optical flow) or image quality features using techniques like Shearlets, integrating them with a multi-layer perceptron. Some composite methods even leverage features extracted from pre-trained deep models. For example, Li et al. [154] employed Principal Component Analysis (PCA) to refine the deep representation obtained from a fine-tuned VGG-Face model. Moreover, methods exploit dynamic features like temporal textures or motion characteristics extracted using LBP-TOP [155] or optical flow [155] [156] from the sequential convolutional representation within the deep learning model. Another approach involves combining deep convolutional features with classical features achieved by fusing

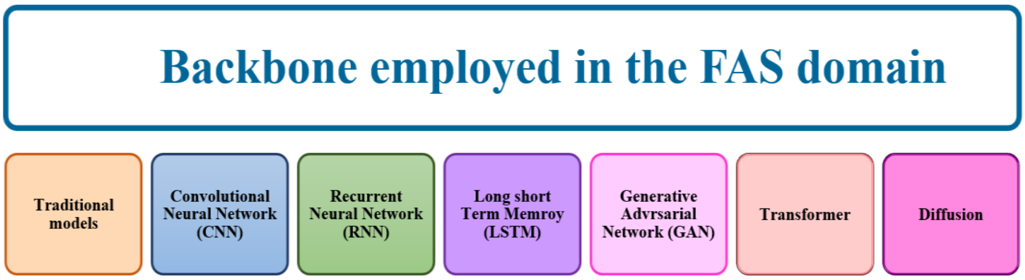

Backbone employed in the FAS domain
Traditional models
Convolutional Neural Network (CNN)
Recurrent Neural Network (RNN)
Long short Term Memroy (LSTM)
Generative Advrsarial Network (GAN)
Transformer
Diffusion

Figure 2.2 Diverse Backbones employed in FAS domain

final scores from both LBP and a deep model like VGG16 [157]. Similarly, Rehmana et al. [158] [159] propose perturbing and modulating low-level convolutional features using HOG [23] and LBP [22] maps to enhance their discriminative power. Li et al. [160] further demonstrate this concept by combining intensity variation features extracted using a 1D-CNN with motion blur characteristics obtained from motion-magnified videos for replay attack detection. LBAS_ResNet50 [161] which combines LBP, ResNet, Bi-directional LSTM, and channel attention mechanism exhibits potential. However, it encounters hurdles related to computational complexity, interpretability, and reliance on blink detection. Lei et al._[162] recent work utilizing a Swin-Transformer architecture for local texture feature extraction with depth information is a promising direction, but further evaluation is needed.

2.3 FAS Backbone architectures

FAS systems have evolved considerably to combat PAs, continuously adapting to overcome challenges, as seen in the preceding sections. Figure 2.2 illustrates the diverse backbones employed in this ongoing journey of innovation. Initially, reliance on traditional features and classifiers led to limited capture accuracy, especially against photo attacks. A pivotal turning point occurred in 2014 when Yang et al. [46] introduced CNNs, demonstrating exceptional feature extraction capabilities for FAS. These data-driven approaches [47] [70] [135]automatically learned intricate hierarchical features from raw data, capturing complex and subtle variation in spoofing patterns. The subsequent integration of deep learning techniques, such as network updating, transfer learning, multi-feature integration, and domain generalization, further propelled the performance of FAS models. However, challenges like the need for extensive annotated data and vulnerability to adversarial attacks prompted the exploration of composite-based methods. These methods integrate traditional features with deep representations, combining the precision of traditional methods with the adaptability of deep learning. The synergy between traditional and deep learning approaches presents an efficient solution, with traditional methods acting as fine-tuners and deep learning techniques providing exponential improvement and adaptability.

FAS systems traditionally analyzed static images, rendering them vulnerable to spoofing attempts using pre-recorded videos or photos. While CNNs excel at extracting features from individual images, they lack the ability to capture the dynamic nature of real faces. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks offer a solution by analyzing facial videos. Both excel at analyzing facial videos, with RNNs adept at capturing sequential information like blinking patterns or expressions that might expose spoofing attempts. However, they struggle with the vanishing gradient problem, limiting their effectiveness in analyzing longer videos. LSTMs, on the other hand, are designed to address this limitation, capturing long-term dependencies, and making them ideal for analyzing longer video sequences in FAS.

As the landscape of neural networks expanded, emergence of Generative Adversarial Networks (GANs) added another dimension of innovation to FAS. GANs, with their dual components—the generator and discriminator—revolutionized the field by training in an adversarial manner. The generator tries to create fake images that look real to effectively deceive the discriminator leading to misclassification of generated samples as real. In this network configuration, the generator benefits from access to discriminator gradients, autonomously guiding its performance enhancement and refining its ability to produce highly realistic images. Lately, works [99] [100] [102] [106] [118] [163] [164] [105] [108] improved the FAS performance indicating that GANs could generate more discriminative images for live and spoof face classification than previous pure CNN methods. Furthermore, the adoption of Transformers in FAS has also marked a significant advancement in the field, departing from the convolutional and recurrent structures of the past. Transformers leverage self-attention mechanisms to capture long-range dependencies among pixels, providing a more comprehensive understanding of the image [86] [87] [165]. These transformer and GAN-based backbones not only enhance the generalization capability of FAS but also achieve state-of-the-art performance on multiple benchmarks. However, these methods are in a progressing phase, and there are drawbacks as well. For instance, Transformers can be computationally intensive and may require more resources compared to CNNs. On the other hand, GANs can be challenging to train due to their involvement in a min-max optimization problem. The quality of the generated images can vary, and poor-quality images may not be beneficial for FAS. Transformer and GAN based methods demands more research.

Besides Transformer and GAN, diffusion models are also emerging as powerful tools in FAS, showcasing impressive capabilities in generative tasks. Double-Diffusion Probabilistic Models (DDPMs) have gained significant attention, adept at learning the intricate patterns of genuine faces while effectively distinguishing spoofed images by gradually introducing Gaussian noise at different scales. However, the effectiveness of these models relies on the initial guide classifier's accuracy in detecting presentation attacks, posing challenges for generalization across diverse spoofing scenarios.

While GANs have demonstrated success in FAS, diffusion models offer certain potential advantages. Generally considered easier to train compared to GANs, diffusion models are less prone to training instability. Additionally, diffusion models provide greater control over the noise injection process, allowing researchers to tailor the training data augmentation for specific types of spoofing attempts.

The choice of backbone architecture for an FAS system hinges on the specific task requirements, considering the unique strengths and weaknesses of each approach. Ongoing research continually refines and pushes the boundaries of FAS technology. Table 2.1 provides a comprehensive overview of backbones used in FAS, detailing their descriptions, advantages, and limitations. This evolution is

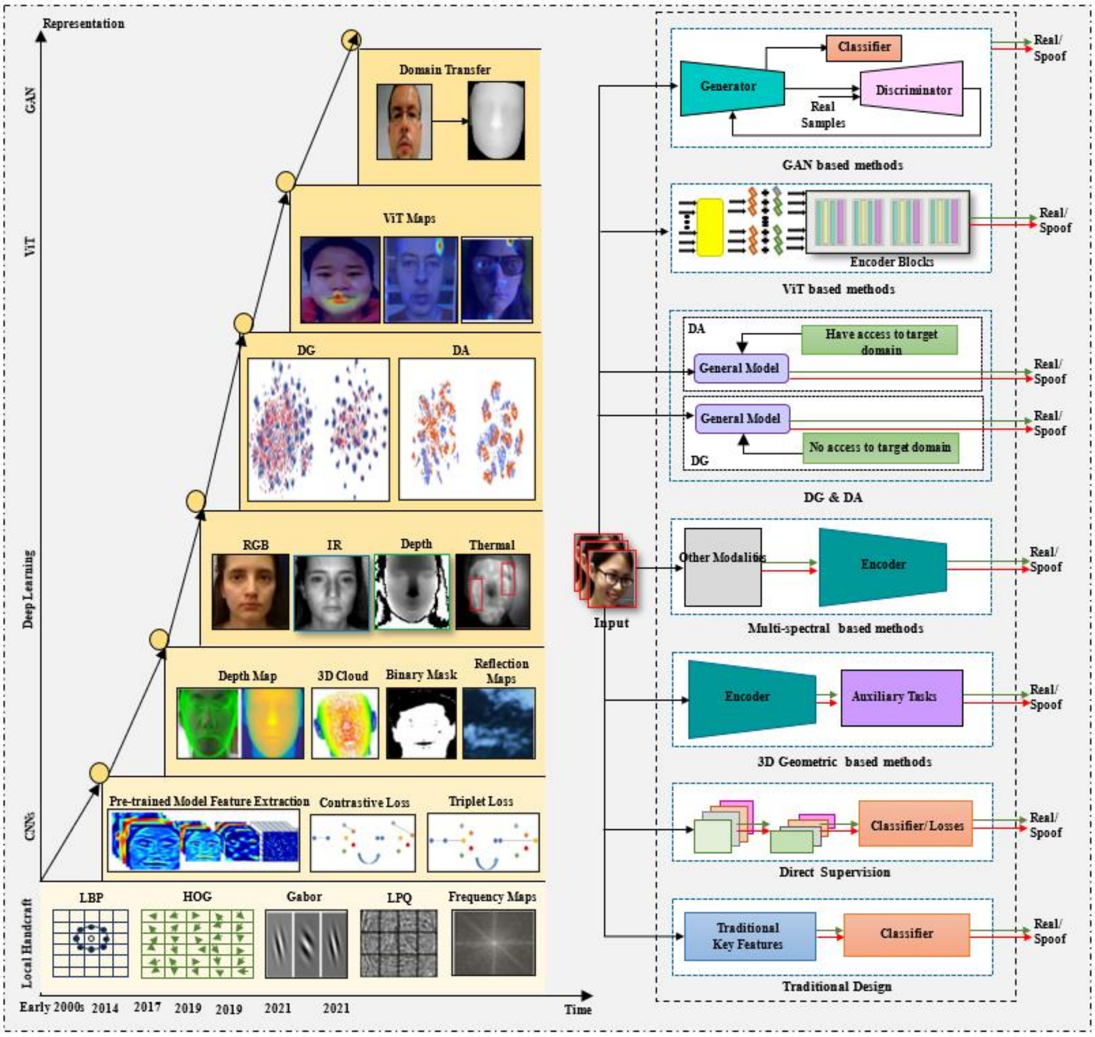


Figure 2.3 Milestone of Face Anti-spoofing frameworks. The rise of handcrafted local descriptors in the early 2000s, followed by the introduction of traditional feature learning approaches in the late 2000s. In 2014, CNN achieved a breakthrough, prompting a shift to deep learning-based strategies such as 3D geometric and multi-spectral methods. The representation pipeline's deepening led to improved performance through Domain Generalization (DG) and Domain Adaptation (DA). Notably, within three years, Vision Transformer (ViT) and Generative Adversarial Network (GAN) integration significantly boosted FAS performance.

also visually depicted in Figure 2.3, which illustrates the progression of backbone structures through milestones, accompanied by their corresponding block diagrams of the general framework. From handcrafted local descriptors in the early 2000s. to the breakthrough of CNNs in 2014, the representation pipeline has

Table 2.1 Comparative Overview of the Backbones employed in FAS

Backbone	Description	Benefit	Limitation
Conventional Models	Employ handcrafted features along with shallow classifiers.	simplicity in implementation and interpretation.	Illumination is sensitive, performs poorly in unconstrained scenarios, and demands a lot of expert knowledge.
Convolutional Neural Network (CNN)	Automatically learn hierarchical features from labelled samples.	Capable of robustly extracting features, particularly adept at capturing spatial information and discerning spoof cues for effective detection.	Susceptible to overfitting, as well as issues with vanishing and exploding gradients.
Recurrent Neural Network (RNN)	Capable of understanding short-term dependencies within a sequence of input samples through learning	Identify temporal inconsistencies across video frames.	Challenging training process; faces issues like vanishing and exploding gradients and struggles with long input sequences.
Long-short Term Memory (LSTM)	Capable of understanding short-term dependencies within a sequence of input samples through learning.	Capable of memorizing features to understand correlations between frames in temporal data.	Susceptible to overfitting, challenging to apply dropout regularization, influenced by varying weight initialization methods.
Generative Adversarial Network	Employs dual neural networks to produce synthetic samples mirroring authentic data distribution closely.	Can be trained to output the pixel-wise likelihood of spoofing attacks for localizing the spoofed regions.	Mode Collapse, Vanishing gradient
Transformers	Utilizes attention mechanisms to capture relationships between different parts of the data.	Excellent at capturing long-range dependencies, potentially useful for subtle spoofing cues.	Limited application in FAS compared to CNNs. Requires more research to fully explore its potential.
Diffusion	A probabilistic model that gradually adds noise to a real image, transforming it into a noise image	Offers a novel approach for learning data distributions and detecting anomalies.	Relatively new technology in FAS, requires further research on effectiveness and computational efficiency.

deepened, with transformers and generative methods achieving a remarkable 99%

performance within just three years. This ongoing progress reflects the commitment to enhancing neural networks for FAS, with each architectural advancement addressing specific challenges and contributing to the effectiveness and efficiency of FAS model.

2.4 Research Gaps

Through an analysis of prior state-of-the-art methods for FAS, several research gaps have been identified as follows:

- Most recent works [56] [166] [167] [168] predominantly concentrate on unimodal inputs (only RGB images) and adopt a single-frame setting (still images). However, it neglects the temporal or multi-modal situations. Thus arises a need for temporal architectures especially for multi modal usage. Generalization and computational cost involved in training the model are two important parameters while designing a framework. As current research focuses heavily on tackling accuracy and generalization issues resulting in increased computation costs thus limiting cost-effective deployment in the practical world.
- Lack of pixel level spoof annotation results in the method being unable to localise spoof regions for visual interpretation. Thus, the model ends up learning unfaithful cues. Therefore, fine-grained pixel-wise spoof segmentation must be advanced to support interpretable FAS and effective auxiliary supervision.
- Need of comprehensive protocols to address the lack of attention towards real-world open-set situations with simultaneous domains and attack types. This will also assist in bridging the gap between academia and industry.
- Use of multimodal protocols in commercially available multimodal devices is cost prohibitive and inefficient as it demands training for each modality such as RGB, RGB-IR, RGB-Depth, RGB-IR-Depth. Thus, indicating a need for adaptive multimodal frameworks which are not constrained to specific modalities.

- Supervised anti-spoofing methods [56] [166] [167] [168] are commonly used in the research community. It is worth exploring semi-supervised and unsupervised anti-spoofing methods that closely deal with real world sample distribution (unlabelled and imbalanced). However, there is high expectation of performance drop in semi-supervised and unsupervised based methods. Therefore, very few works have been observed in the literature.

- Rotation, horizontal, and vertical flip contrast variation are some of the commonly used data augmentation techniques. However, employing adversarial learning could offer an effective means of adaptive data augmentation, particularly when targeting broader and more heterogeneous domains.

CHAPTER 3

TWO-STREAM RGB-BASED FAS FRAMEWORK

Early face analysis methods used global representations [8] [9], treating the entire facial image as a single vector and applying dimensionality reduction techniques like PCA [10] and LDA [11]. Although effective in controlled environments, these approaches often failed under real-world conditions involving pose variations, illumination changes, and occlusions. To enhance robustness, researchers turned to local feature descriptors—such as POEM [12], HGPP [13], LBP [14], and LGBP [15] —which proved more robust in uncontrolled settings. Recent works have further advanced the field by integrating handcrafted texture features with deep learning techniques [16], enabling stronger generalization across diverse environments.

Building on these advancements, this chapter introduces a novel two stream FAS framework based on RGB input, designed to address challenges such as illumination variability and environmental inconsistencies. The framework is rigorously evaluated through systematic preprocessing, detailed experimental setups, comparative analysis, ablation studies, and comprehensive discussion.

3.1 Methodology

When two distinct branches are employed in parallel to process complementary feature representations and are later fused for joint decision-making, the resulting design is referred to as a two-stream architecture. In the context of the proposed framework, this architecture is termed a two-stream RGB-based framework, as it leverages two parallel branches that operate on RGB facial inputs in different ways as shown in Figure 3.1.

The first branch utilizes a modified Xception network to extract deep semantic features directly from raw RGB images. The second branch focuses on fine-grained texture analysis by applying a Multi-level Elliptical Local Binary Pattern (ELBP) descriptor to face regions represented in alternate color spaces (YCbCr and HSV). The fusion of deep and texture features enables the framework to capture both global and

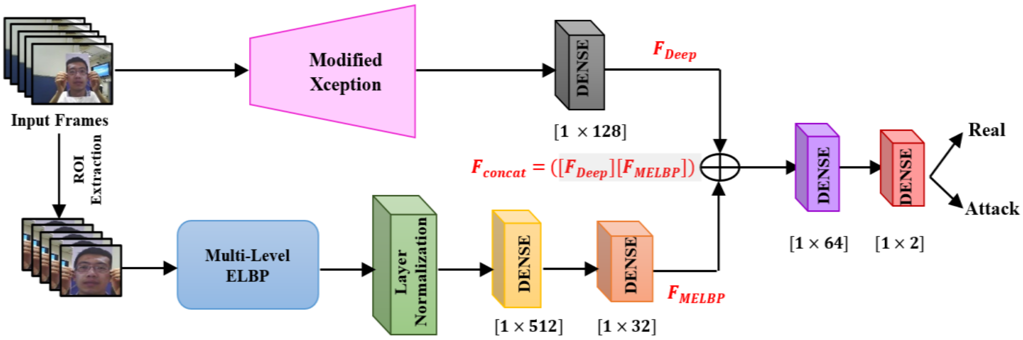


Figure 3.1 Proposed two stream FAS framework with multi-level ELBP and modified Xception.

local facial cues, enhancing its robustness against spoofing under varied illumination and environmental conditions. The following sections describe each component of the framework in detail.

3.1.1 Pre-processing step

The datasets comprise videos of both genuine and spoofed faces. During preprocessing, every 10th frame is uniformly sampled, labeled as per dataset protocols, and resized to $[299 \times 299]$ pixels for input into the modified Xception network. Face detection and region of interest (ROI) extraction are initially performed using the Viola-Jones algorithm [169]. Due to localization failures (4.3% in Replay-Attack, 3.7% in CASIA, and 7% in MSU-MFSD), MTCNN [170] is used as a fallback. The extracted facial regions, along with a margin of background pixels, are resized to $[150 \times 150]$ pixels. Further, we randomly flip the input data (both raw and ROI data) horizontally.

3.1.2 Color space

RGB is the most widely used color space in digital imaging for representation, sensing, and display. While effective for general image depiction and visual perception, RGB often falls short in capturing subtle artifacts and high-frequency textural variations that are crucial for face anti-spoofing tasks. Recent studies [171] [172] have highlighted the limitations of relying solely on RGB for spoof detection, as it struggles to distinctly capture spoof-related artifacts such as print defects, screen noise, moiré patterns, and illumination inconsistencies. This is due to the mixing of luminance and chrominance information in the RGB space. In contrast, alternative color models like HSV (Hue, Saturation, Value) and YCbCr (Luminance, Chrominance Blue, Chrominance Red) offer a more structured representation by separating intensity and color information. For example, the HSV color space isolates chromatic content (hue and saturation) from intensity (value), making it more robust to lighting variations and better at highlighting color-based spoofing cues. Similarly, YCbCr separates luminance (Y) from chrominance (Cb and Cr), which is particularly useful for capturing texture distortions and illumination-related inconsistencies—common indicators of spoof attempts. Incorporating these color spaces allows the model to exploit complementary feature representations that may not be evident in the RGB domain alone.

3.1.3 Feature extraction

Texture and deep features are independently extracted through two separate branches, as detailed in the following sections.

3.1.3.1 ELBP based texture features

Zhao et al. [173] identified the eyes and mouth—regions with elliptical structures—as key areas for facial analysis. Leveraging this, the Elliptical Local Binary Pattern (ELBP) descriptor is adopted as an effective tool for extracting micro-textural and gradient features, outperforming conventional LBP.

The ELBP formulation considers each pixel $(X_p\,, Y_p)$ as the center of an ellipse defined by its semi-major and semi-minor axes R_1 and R_2, respectively. The descriptor encodes the local texture by sampling pixel intensities along the elliptical perimeter and comparing them to the center pixel. The ELBP code for a given configuration is expressed as:

$$ELBP^{N,R1,R2}\,(X_p\,, Y_p) = \sum_{i=1}^{N} F(g_i^{N,R1,R2} \text{-} \ g_p)2^{i-1} \tag{3.1}$$

where the thresholding function F(x) is defined as:

$$F(x) = \begin{cases} 0, \ if \ \ x < 0 \\ 1, \ if \ \ x \geq 0 \end{cases} \tag{3.2}$$

Neighbour coordinates are determined via angular steps $\Delta\Theta \ = 2 \times \pi/N$, yielding the coordinates:

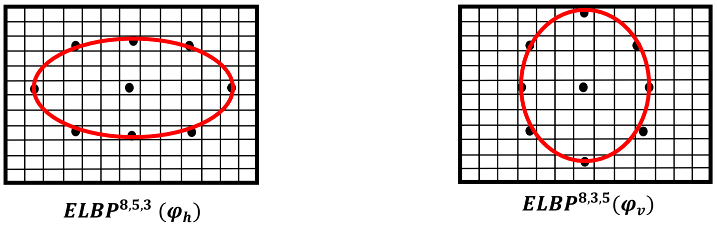


Figure 3.2 Horizontal and Vertical ELBP Pattern

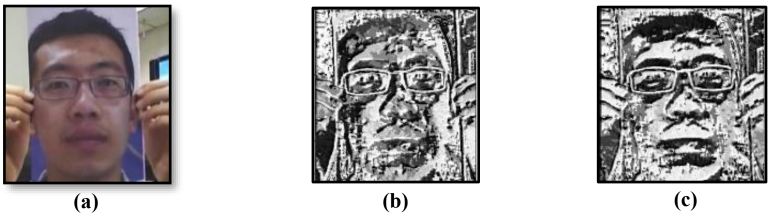


Figure 3.3 An original image (a), $ELBP^{8,5,3}$ (b), $ELBP^{8,3,5}$ (c)

$$x_i = x_p + R1 \times cos\ ((i-1) \times \Delta\Theta)) \tag{3.3}$$

$$y_i = y_p - R2 \times \sin\ ((i-1) \times \Delta\Theta)) \tag{3.4}$$

To construct the ELBP feature vector, the entire ELBP image is first divided into non-overlapping sub-regions. For each sub-region, a histogram of ELBP codes is computed and subsequently concatenated to form the final representation. Based on ellipse shape, the operator φ is categorized as:

$$\varphi = \begin{cases} LBP, & R1 = R2 \\ \varphi_v, & R1 < R2 \\ \varphi_h, & R1 > R2 \end{cases} \tag{3.5}$$

Inspired by [174], both vertical ($\varphi_v = ELBP_v,= ELBP^{8,R1,R2}$) and horizontal $\varphi_h = ELBP_h = ELBP^{8,R2,R1}$) are used to enrich the descriptor as shown in Figure 3.2 and 3.3. Their respective feature vectors are concatenated, enriching the descriptor's discriminative power.

For a given input frame, the image is segmented into a grid of $W \times H$ subregions. When both vertical and horizontal patterns are applied, the resulting ELBP feature vector has a dimensionality of $2 \times W \times H \times 256$. However, to reduce computational load and memory footprint, uniform patterns are employed—restricting the number of distinct patterns to 59. This optimization reduces the vector length to $2 \times W \times H \times 59$, respectively, making the approach more tractable for large-scale or real-time applications.

To further enhance feature granularity, multi-level segmentation is applied.

Algorithm 1: Algorithm of multi-level ELBP Feature Vector ($\mathcal{F}$)

```
Input:   Video Frame F, Color Space C, ELBP parameters ( $8, R_1, R_2$)
Output: Feature vector H of Frame F
$I$ ← Face Detection and Cropping ($F$)
$I^i$ ← Color Space Conversion ($I, C$)        # $N$ is the number of color channels in $C$
$L_0$ ← 0, $L_1$ ← 1, $L_3$ ←2        # $L_0$,$L_1$ is the first, second and third level of sub-region divisions, respectively
for $i$ ← 1 to N do
    $I_C^i$ ← ELBP ($I^i, 8, R_1, R_2$)
    $H_C^i(L_0)$ ←uniform histogram ($I_C^i$)        #statistic the histogram of $I_C^i$
    for j ← 1 to 4 do
        h($j$) ← ELBP ($I^i$ ($j$), $8, R_1, R_2$)        # ELBP of each sub region to histogram
        $H_C^i(L_1)$ ← append $h(j)$        # append $h(j)$ to the end of $H_C^i(L_1)$
    end for
    for k ← 1 to 16 do
        h($k$) ← ELBP ($I^i$ ($k$), $8, R_1, R_2$)        # ELBP of each sub region to histogram
        $H_C^i(L_2)$ ← append $h(k)$
    end for
end for
$H_C^i$ = concatenate ($H_C^i(L_0), H_C^i(L_1) H_C^i(L_2)$)        # concatenate $H_C^i(L_0)$ , $H_C^i(L_1) H_C^i(L_2)$
$H_C$ = append $H_C^i$
end for
```

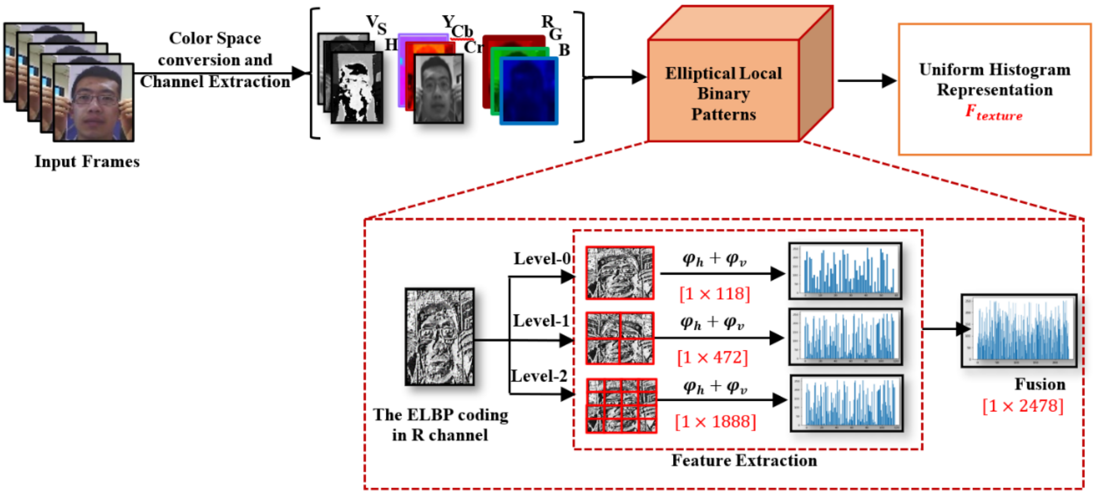


Figure 3.4 Multi-Level ELBP based texture feature extraction.

The face region of interest (ROI) is recursively divided into 4^l sub-regions for levels l ϵ {0,1,2,3}. Beyond L_2, sub-regions become too small (e.g., 18 × 18), leading to high-dimensional and noisy features, with no performance gains.

The histogram for each sub-region is computed as:

$$H(n)\text{=}H(n) = \sum_{i=0}^{X} \sum_{j=0}^{Y} \varphi^{N,R_1,R_2}(i,j) \tag{3.6}$$

For a given color space C with k channels, the complete ELBP feature vector:

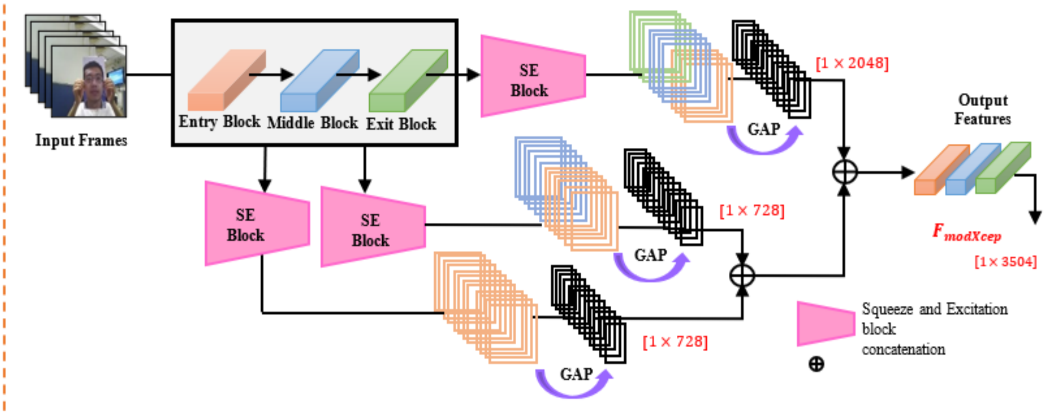


Figure 3.5(a) Modified Xception Network

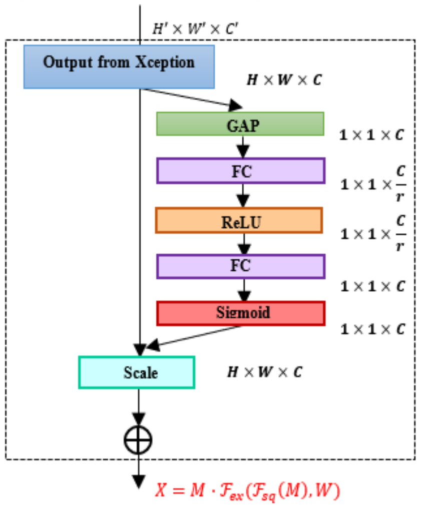


Figure 3.5(b) Xception [175] with integrated Squeeze and Excitation (SE) Block [2]

$$H_c = [H_c^1 H_c^2 \ldots\ldots H_c^k] \tag{3.7}$$

The complete ELBP extraction, including level-wise processing, is detailed in Algorithm 1. Figure 3.4 visualizes feature extraction across levels 0 to 2.

3.1.3.2 Modified Xception based deep features

In the framework, the Xception network [175] is used as the baseline for deep feature extraction due to its efficient architecture, which separates channel and spatial correlation learning. Its three-stage structure—entry, middle, and exit flows—offers high accuracy with fewer parameters, making it well-suited for face anti-spoofing. To enhance the model's resilience against unseen attack types, we integrate Squeeze-and-Excitation (SE) [2] modules into each of the three flows, as illustrated in Figure 3.5(a).

An SE block, shown in Figure 3.5(b), improves channel-wise feature representation through two stages: **squeeze** and **excitation.** Let the transformation $\mathcal{F}_{tr}$, implemented via Xception blocks, map an input image $I\epsilon\mathbb{R}^{H'\times W'\times C'}$ to a deep feature map $M\epsilon\mathbb{R}^{H\times W\times C}$. The SE block processes this feature map in two steps:

Squeeze: Spatial information is aggregated through global average pooling, yielding a compact channel descriptor:

$$z_C = \mathcal{F}_{sq}(m_C) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} m_C(i,j) \tag{3.8}$$

Excitation: This descriptor is passed through a gating mechanism, which learns inter-channel dependencies using two fully connected layers and a sigmoid activation:

$$s = \mathcal{F}_{ex}(z,W) = \sigma\big(g(z,W)\big) = \sigma(W_2\delta(W_1z)) \tag{3.9}$$

here δ denotes the ReLU activation function [176] weight matrices $W_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $W_2 \in \mathbb{R}^{C\times\frac{C}{r}}$ and r is the reduction ratio, typically used to control model complexity.

The recalibrated output X is obtained by applying these weights to the feature mappings M:

$$X = \mathcal{F}_{scale}(m_C, s_C) = s_C m_C \tag{3.10}$$

This SE block can be seamlessly integrated into standard CNNs by placing it after the non-linearity that follows each convolution. It provides a favorable balance between performance improvement and computational overhead.

In our framework, SE modules are embedded into the pre-trained Xception model at the entry, middle, and exit flows. The resulting feature maps have dimensions of $[19\ \times\ 19\ \times 728]$ for both the entry and middle flows, while the exit flow produces a feature map of size $[10\ \times 10 \times 2048]$. To aggregate these representations, we apply global average pooling (GAP) independently to the output of each SE-enhanced flow. The resulting vectors are then concatenated to form a unified and compact deep feature representation, capturing both spatial and channel-specific information across multiple levels of abstraction.

3.1.4 Classification model

After feature extraction, five dense blocks $D_i, i \in (1,5)$ are employed to learn discriminative patterns for classifying real and spoof faces. The respective dimensions of D_1, D_2, D_3, D_4 and D_5 are $([1 \times 128], [1 \times 512], [1 \times 32], [1 \times 64]$ and$[1 \times 32]$ respectively. The dense blocks operate on feature sets from two branches: F_{Deep}, obtained from the Modified Xception Deep branch, and F_{MELBP}, derived from the Multilevel ELBP branch. While F_{Deep} captures semantic features from the full image, F_{MELBP}, encodes multi-scale texture patterns from cropped facial regions. In the ELBP stream, D_2 and D_3 introduce non-linearity and convert handcrafted texture descriptors into learnable vectors.

Once the features are processed and dimensionally reduced, they are concatenated as $F_{concat}([F_{Deep}][F_{MELBP}])$. This combined representation is passed through D_4 to enhance joint feature learning before classification via D_5 which uses a sigmoid activation. All preceding dense layers employ ReLU activation [176].

3.2 Database and Experimental Analysis

The following section outlines the datasets, experimental methodology, and ablation studies conducted to evaluate the proposed framework and assess the impact of key design choices.

3.2.1 Dataset

To evaluate the robustness and effectiveness of the proposed work, experiments were conducted on three widely used FAS dataset. A detailed description of each dataset is provided below:

CASIA-FASD: It consists of 600 video clips from 50 subjects, captured using multiple imaging devices under three distinct image quality settings: low, normal, and high. The dataset is divided into a training set comprising 240 videos from 20 subjects and a testing set of 360 videos from the remaining 30 subjects. To reduce computational complexity and storage demands, high-quality videos were downsampled to a resolution of 1280 × 720 pixels. The dataset includes three types of spoofing attacks: cut photo attacks, warped photo attacks, and video replay attacks.

REPLAY-ATTACK: This dataset includes 1,200 short video clips from 50 individuals, captured under varying illumination conditions (controlled and adverse) and two support setups (handheld and fixed). It comprises 200 genuine access attempts, 400 mobile-based replay attacks, 200 printed photo attacks, and 400 high-resolution screen replay attacks. Although the dataset is structured into four folders, only three subsets—train, validation, and test—were utilized in this study. The attack modalities include:

(i) mobile screen attacks using an iPhone 3GS, (ii) high-resolution replay attacks using an iPad, and (iii) print attacks using A4-sized photographs.

MSU-MFSD: This database was developed with the participation of 35 subjects. Each participant contributed two real access videos recorded using Android smartphones and laptop webcams. High-definition videos were captured using both a Canon DSLR and an iPhone. The iPhone recordings were replayed on mobile screens to simulate mobile replay attacks, while Canon footage was displayed on an iPad Air to generate high-definition screen replay attacks. For photo attacks, facial images of all participants were printed on A3-sized paper using an HP color printer. The dataset was divided into two subsets: a training set with 120 videos from 15 subjects, and a testing set with 160 videos from the remaining 20 subjects.

3.2.2 Evaluation Metrics

For comparative evaluation against existing state-of-the-art methods, the performance of the proposed model is assessed using standard metrics defined as follow:

Equal Error Rate (EER): It represents the operational threshold where False Acceptance Rate (FAR) equals False Rejection Rate (FRR) i.e.,

$$FAR = FRR \tag{3.11}$$

Half Total Error Rate (HTER): It provides a threshold-dependent performance measure calculated as:

$$HTER = \frac{FRR+FAR}{2} \tag{3.12}$$

In addition to these metrics, the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) are also reported to provide a comprehensive evaluation of the model's discriminative capability across various threshold settings.

3.2.3 Implementation details

The framework was implemented using the Keras, and all the experiments were conducted on Google Colab Pro utilizing an NVIDIA P100 GPU with 24 GB of RAM. For image pre-processing, video data were sampled at every 10th frame, and horizontal flipping was applied as a data augmentation technique to enhance model generalization. The network was trained using the Adam optimizer with a learning rate

Table 3.1 Intra-dataset testing: CASIA FASD, Replay-Attack and MSU-MFSD datasets (%).

Methods	**CASIA FASD**		**Replay-Attack**		**MSU-MFSD**	
	EER	**HTER**	**EER**	**HTER**	**EER**	**HTER**
YCbCr + HSV-LBP [22]	6.2	-	**0.4**	2.9	-	-
DMD [177]	21.7	-	5.3	3.7	-	-
CNN LBP+TOP [155]	8.02	9.94	3.22	4.70	-	-
Fine-tuned VGG- Face [154]	5.2	-	8.40	4.3	-	-
DPCNN [154]	4.5	-	2.90	6.1	-	-
Colour SURF [26]	2.8	-	0.1	2.2	2.2	
Hybrid CNN [178]	**2.2**	-	0.5	1.6	-	-
RI-LBP + Speeded Up Robust Features (SURF) [179]	4.6	-	1.2	4.2	1.5	
Colour Texture Markov Feature (CTMF) [172]	8.0	-	4.0	4.4	7.5	
LSCNN [52]	4.44	-	0.33	2.50	-	-
RI-LBP + Deep Features [180]	4.4	-	2.3	2.6	3.1	-
Adversarial [112]	3.2	-	**-**	1.4		
Patch CNN [64]	4.8	-	1.5	-	-	-
Proposed method ($\boldsymbol{L_0 + L_1}$)	6	8.7	0.5	0	0.10	0.04
Proposed method ($\boldsymbol{L_0 + L_1 + L_2}$)	**2.37**	**3.2**	**0**	**0**	**0**	**0.06**

set to $10e - 4$. A batch size of 32 was employed, along with random shuffling and flipping of training images to introduce variability during training. Both branches were independently trained on each dataset. Upon completion of individual training, both branches were frozen for subsequent adaptation and integration in the final model.

3.2.4 Comparative Analysis with other-state-of-the-arts

This section presents the evaluation of the proposed framework using both intra-dataset and inter-dataset testing on three benchmark datasets*:* CASIA FASD [7], REPLAY-ATTACK [181], MSU-MFSD [8]. Performance is reported using EER and HTER, as shown in Tables 3.1 and 3.2.

3.2.4.1 Intra-dataset testing

In intra-dataset testing, training, and testing are performed on the same dataset, allowing assessment of the model's ability to learn from and classify within a controlled environment. The proposed framework achieves superior performance on all three datasets as shown in Table 3.1. Notably, it records 0% EER and HTER on the Replay-Attack and 0% HTER, 0.06% EER on MSU-MFSD, indicating high accuracy

Table 3.2 Cross-dataset evaluation on CASIA FASD, Replay-Attack and MSU-MFSD in terms of HTER (%).

Train/Test Datasets **Methods**	**CASIA FASD**		**Replay-Attack**		**MSU-MFSD**	
	Replay-Attack	**MSU-MFSD**	**CASIA FASD**	**MSU-MFSD**	**CASIA FASD**	**Replay-Attack**
Colour SURF [26]	26.9	31.8	**23.2**	**19.1**	24.3	29.7
RI-LBP + SURF [179]	**9.6**	19.8	39.2	33.3	29.7	21.4
CTMF [172]	32.3	32.4	45.9	37.7	57.0	42.7
RI-LBP+ Deep Features [180]	35.3	**2.1**	38.5	20.6	32.4	35.8
Adversarial [112]	17.5	**9.3**	41.6	30.5	**17.7**	**5.1**
DRL-FAS [56]	28.4	-	33.2	**15.6**	-	29.7
CFSA-FAS [182]	24.3	-	34.0	25.2	-	22.8
GFA-CNN [117]	21.3	-	34.3	23.5	-	25.8
SAPLC [96]	27.31	-	37.50	-	-	-
CDCN++ [70]	6.5	-	29.8	-	-	-
DR-UDA [113]	**15.6**	**9.0**	34.2	29.0	**16.8**	**3.0**
Proposed method ($L_0 + L_1$)	44.71	30.6	36.89	35.37	35.27	37.7
Proposed method ($L_0 + L_1 + L_2$)	**20.05**	**9.5**	**28**	**21.8**	**23.02**	**24.3**

and strong adaptability to the spoofing types present in these datasets. For the more challenging CASIA-FASD dataset, ELBP segmentation up to Level L_2 reduces both EER and HTER by 50%, achieving 2.37% and 3.2%, respectively. These results confirm the model's robustness in controlled settings. However, to evaluate its real-world applicability, inter-dataset testing is also conducted to assess cross-domain generalization.

3.2.4.2 Inter-dataset testing

Inter-dataset testing assesses the model's robustness against unseen spoofing attacks by training and testing on different datasets. Each configuration was evaluated three times, with average results reported in Table 3.2:

- **Test Case 1** (**Train**: CASIA-FASD, **Test**: REPLAY-ATTACK & MSU-MFSD): The framework achieves HTER of 20.05% on Replay and 9.5% on MSU. While prior methods [70] [179] have reported lower HTERs on specific combinations, their performance often lacks consistency across datasets. In contrast, our model demonstrates more stable and reliable generalization across different testing environments.

- **Test Case 2** (**Train:** REPLAY-ATTACK, **Test***:* CASIA-FASD & MSU-MFSD): Higher HTERs are observed in this setting—28% for CASIA and 21.8% for MSU—primarily due to the limited variability in the Replay dataset compared to CASIA. Despite these challenges, the proposed framework outperforms most existing methods, except [26] [56]**.** These results underscore the importance of training data diversity for effective cross- dataset generalization.
- **Test Case 3** (**Train:** MSU-MFS, **Test:** CASIA-FASD & REPLAY-ATTACK): The proposed model achieves HTER of 23.02% on CASIA and 24.3% on Replay. Although these results are lower than intra-dataset performance, they still outperform many existing methods and demonstrate improved stability and generalization across diverse datasets.

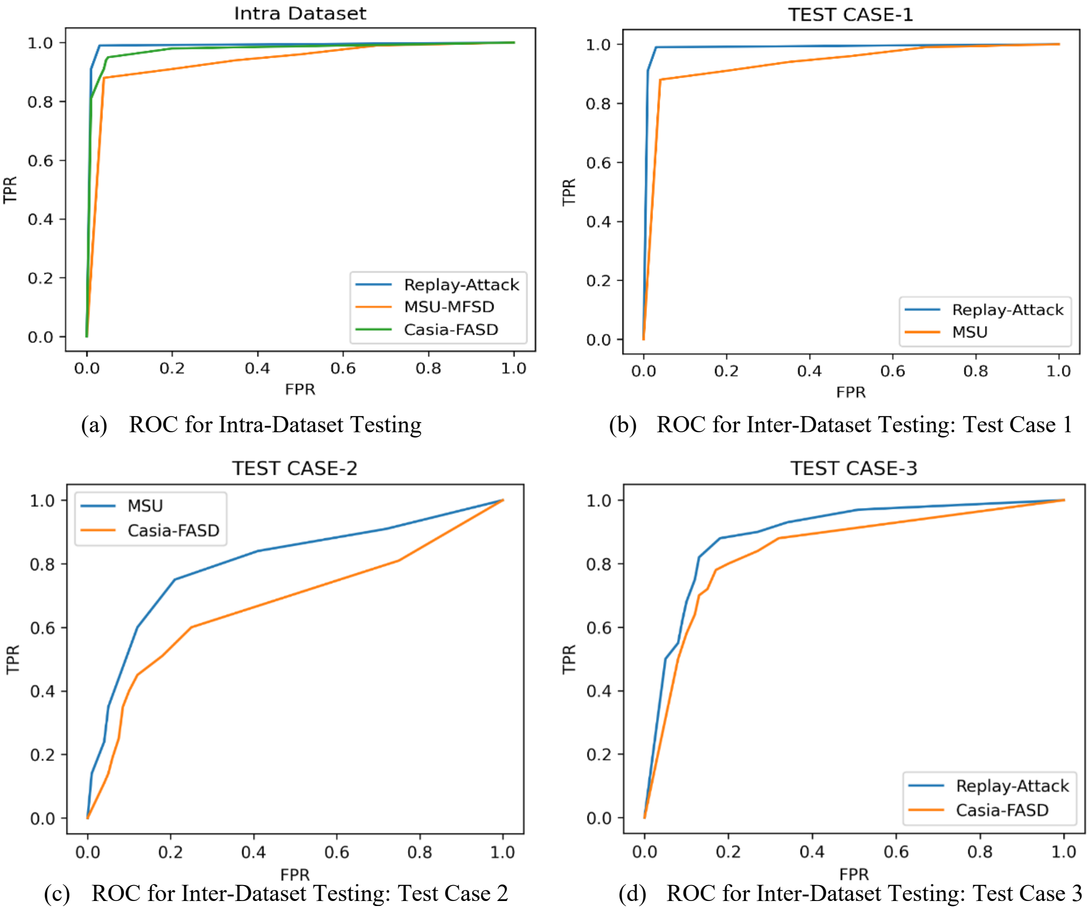
Unlike prior works with inconsistent results across datasets, our framework maintains HTERs below 30% in all inter-dataset tests and under 4% in intra-dataset settings. This highlights its robustness and unbiased performance across varied conditions.

To provide a more visual comparison, Figure 3.6 presents ROC curves demonstrating consistently high true positive rates across evaluation scenarios.

3.2.5 Ablation Analysis

To validate the effectiveness of our framework, a series of ablation studies were conducted, focusing on baseline model selection, architectural variations, color space impact, dense layer configuration, and runtime performance.

Analysis for selection of baseline: The Xception model, though highly effective in various computer vision tasks, remains underexplored in FAS. Motivated by its potential, it was selected as the baseline for this study. A comparative analysis with VGG16 [183], ResNet50 [184], Xception [175], EfficientNet [185], and Inception [186]—evaluating accuracy, trainable parameters, and model size (Figure 3.7)—confirmed this choice. While ResNet50 performed well, it required significantly more parameters. In contrast, the modified Xception model delivered superior accuracy with



(a) ROC for Intra-Dataset Testing

(b) ROC for Inter-Dataset Testing: Test Case 1

(c) ROC for Inter-Dataset Testing: Test Case 2

(d) ROC for Inter-Dataset Testing: Test Case 3

Figure 3.6 ROC curves for (a) intra-dataset testing; (b) Inter-dataset testing: Test case 1- Trained on CASIA-FASD, Tested on Replay-Attack and MSU-MFSD, (c) Inter-dataset testing: Test Case 2- Trained on Replay-Attack, Tested on CASIA FASD and MSU-MFSD, (d) Inter-dataset testing: Test case 3- Trained on MSU-MFSD, Tested on CASIA- FASD and Replay-Attack.

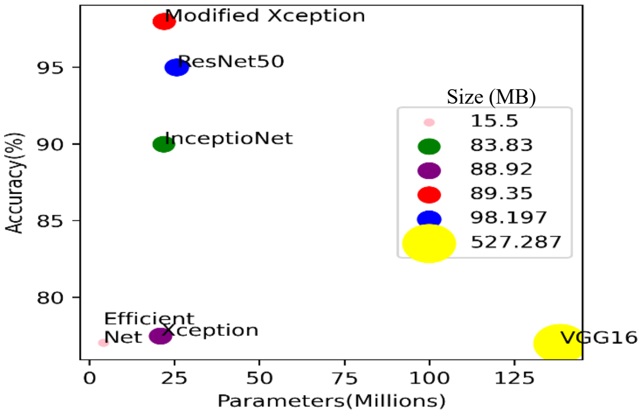


Figure 3.7 Comparison of different generic architectures for baseline selection.

Table 3.3 Ablation Study: Intra-dataset evaluation of modified Xception architecture on Reply-Attack, CASIA-FASD, and MSU-MFSD (%).

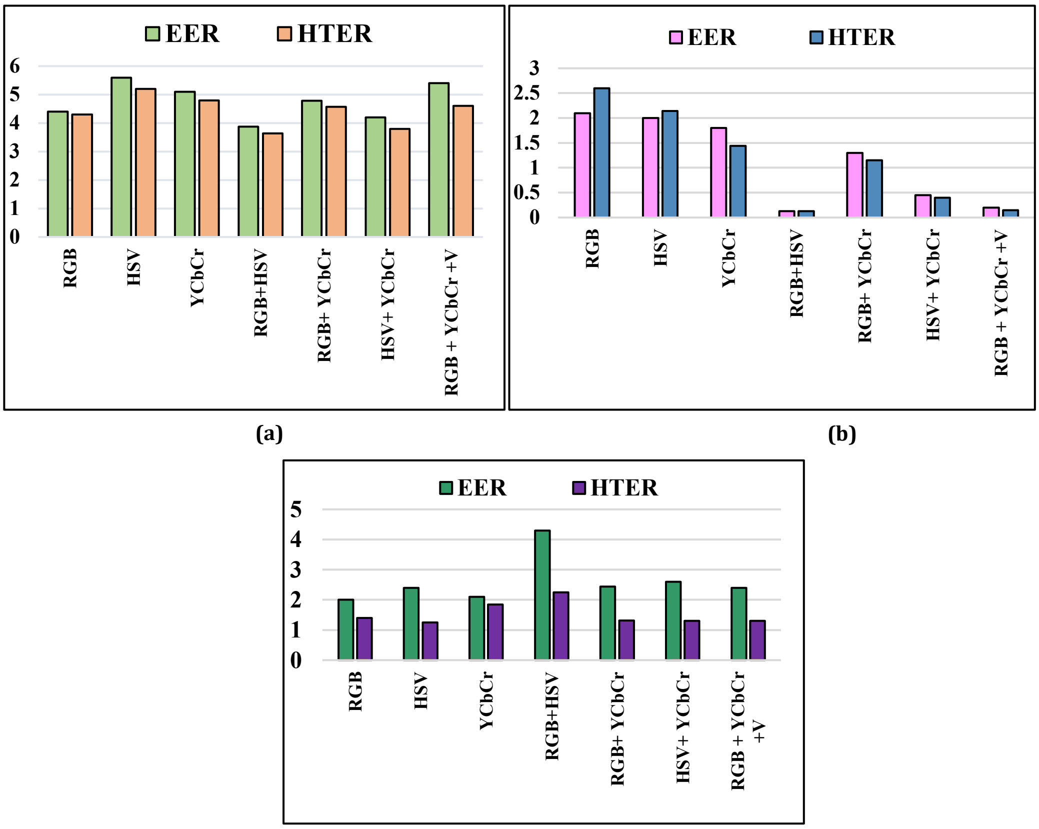
Variation	CASIA FASD		Replay-Attack		MSU-MFSD	
	EER	HTER	EER	HTER	EER	HTER
Baseline Xception **[175]**	50.1	50	36.73	29.03	33.11	25.29
SE at Entry level	14	10.5	3.64	1.89	10.68	9.7
SE at Middle Level	11.3	9.1	8.28	4.68	8.08	4.5
Modified Xception	**7.45**	**5.08**	**2.45**	**1.5**	**0.92**	**0.46**

minimal added complexity, making it the most efficient and suitable baseline.

Analysis of variation in baseline Xception model: We introduced Squeeze-and-Excitation (SE) modules after each block (entry, middle, and exit) in the original Xception model (Figure 3.5(a)) to improve multi-level feature representation. The modified model, trained for 10 epochs across all three datasets, significantly reduced EER and HTER compared to the baseline (Table 3.3), demonstrating its superiority in extracting deep features.

Analysis of the effect of different color space: RGB encodes images using primary colors but lacks perceptual fidelity, whereas HSV and YCbCr separate color and luminance components, enhancing feature representation. The modified Xception branch accepts tri-channel inputs, while the ELBP branch extracts features per channel across RGB, HSV, and YCbCr. Extensive evaluations under intra- and inter-dataset testing (Figure 3.8) show that RGB+HSV performs best on CASIA and MSU-MFSD, while RGB+HSV+YCbCr yields the lowest HTER on Replay-Attack. However, in inter-dataset settings, RGB and HSV alone perform poorly, confirming that RGB+HSV+YCbCr offers superior generalization (Table 3.4). This combination, using Level 0 ELBP, consistently outperforms others, and as illustrated in Figure 3.9, multi-level ELBP fusion (L_0+L_1+L_2) with modified Xception features significantly enhances performance.

Ablation Study on Dense Layer Configuration: To enable binary classification, the feature vectors from the modified Xception model, $F_{modXcept}$ $[1 \times 3504$ (Fig. 5(a)) and multi-level ELBP features, $F_{texture}$ $[1 \times 2478]$are passed through a series of



(a)

(b)

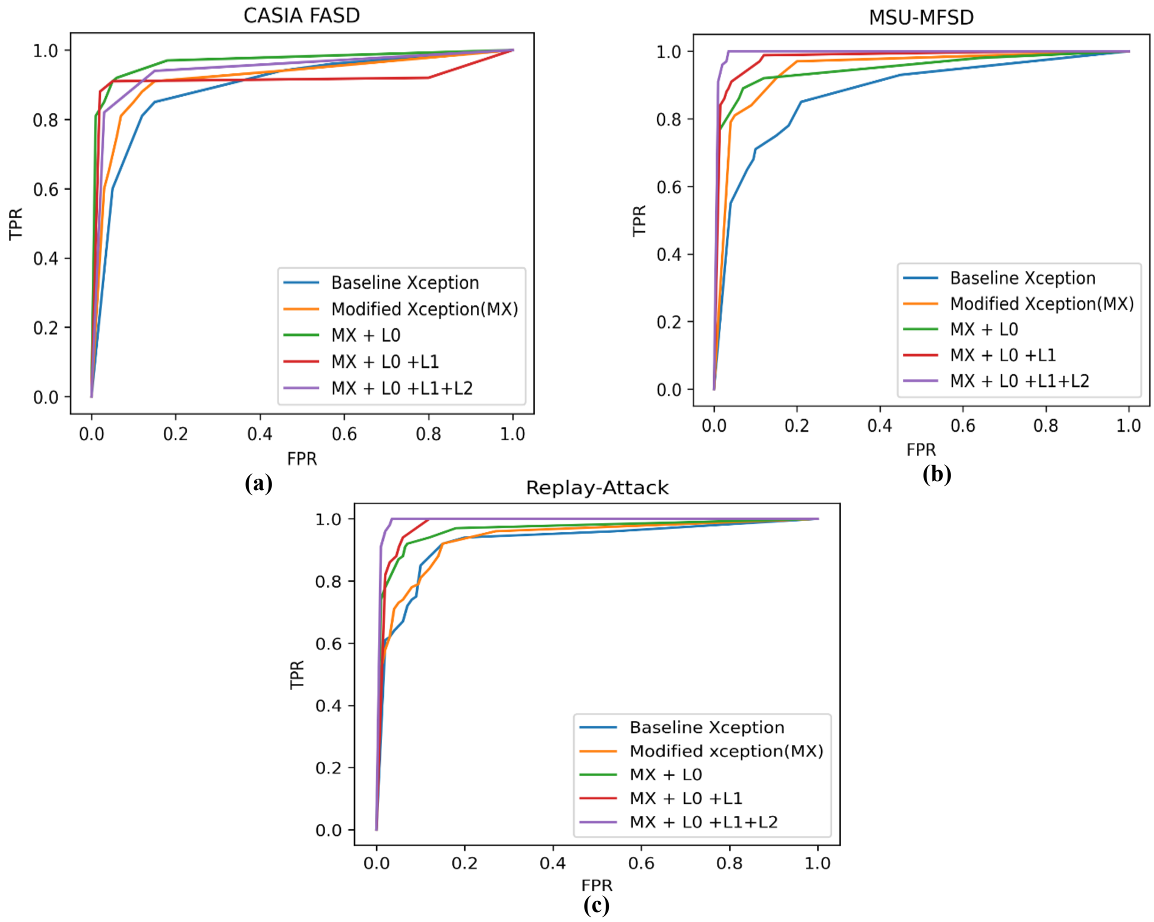
(c)

Figure 3.8 Intra-dataset testing EER and HTER obtained on (a) CASIA FASD (b) MSU-MFSD (c) Replay-Attack.

Table 3.4 Cross-dataset evaluation of Level-0 ELBP across different color spaces on CASIA-FASD, MSU-MFSD and Reply-Attack for ablation analysis (HTER %).

Train	**CASIA FASD**		**MSU-MFSD**		**Replay-Attack**	
Test	**MSU MFSD**	**Replay-Attack**	**CASIA FASD**	**Replay-Attack**	**CASIA FASD**	**MSU MFSD**
RGB	27.88	30.35	30.55	29.87	23.53	22.75
HSV	27.69	28.77	31.68	29.51	24.39	22.14
YCbCr	26.08	31.20	30.04	26.70	17.76	16.29
RGB + HSV	27.78	31.08	30.18	29.99	24.75	21.78
RGB + YCbCr	27.49	28.86	30.37	29.20	23.92	22.31
HSV+ YCbCr	22.35	26.1	21.5	25.6	24.54	22.17
RGB + YCbCr +HSV	**18.61**	**20.41**	**18.21**	**19.89**	**22.69**	**18.79**

dense layers. Following extensive experimentation (Table 3.5), the optimal configuration was found to be $[128, 512, 32, 64, 2]$ delivering the best results across all datasets.



(a)

(b)

(c)

Figure 3.9 Performance of the proposed model with different designs on (a) CASIA FASD (b) MSU MFSD (c) Replay Attack using Intra-dataset testing.

Table 3.5 Selection of Nodes in Dense Layers (D_i)

D_1	D_2	D_3	D_4	D_5	Accuracy			HTER		
					CASIA	**RA**	**MSU**	**CASIA**	**RA**	**MSU**
256	256	-	64	2	96.354	98.21	98.24	3.7	0.02	0.84
256	512	64	32	2	94.44	99	99.54	4.5	0	0.78
128	256	32	64	2	97.6	99.54	98.35	3.55	0.02	1.2
128	512	-	32	2	98.1	98.54	99.84	3.1	0.01	0.7
128	**512**	**32**	**64**	**2**	**98.32**	**99.98**	**99.76**	**3.2**	**0**	**0.06**

Model Runtime: The modified Xception stream alone requires ~0.71 seconds per batch of 32 images, while the complete dual-stream framework takes ~0.82 seconds due to additional face cropping and ELBP computations. This results in an effective throughput of approximately 40 frames per second.

3.3 Significant Outcomes

The outcomes of this chapter are summarized as follows:

- A two-stream, multi-level face anti-spoofing framework is introduced, combining multi-level ELBP texture features with modified Xception-based deep features for efficient spoof detection.
- To mitigate overfitting, the framework utilizes a modified Xception network with squeeze-and-excitation mechanisms, allowing for the extraction of multi-level deep features without increasing the number of parameters. The multi-level ELBP descriptor captures textural details while ensuring a balanced trade-off between performance and parameter efficiency.
- Comprehensive ablation studies demonstrate the effectiveness of multi-level texture feature representation across various color spaces, confirming the framework's ability to enhance deep feature extraction while maintaining optimal performance.
- While the framework primarily uses the RGB modality, integrating additional modalities could improve robustness.
- A key challenge lies in effectively fusing multiple modalities without significantly increasing model complexity. Chapter 4 will explore a framework that incorporates other modalities, balancing complexity, and performance.

CHAPTER 4

MULTIMODAL VISION TRANSFORMER FOR ROBUST FACE ANTI-SPOOFING

Relying solely on a single modality (e.g., RGB) significantly limits the effectiveness of FAS systems in detecting increasingly sophisticated PAs, particularly as the quality and realism of PAIs continue to advance. The rapid evolution of spoofing techniques has rendered single-modal systems inadequate in capturing the diverse and subtle spoofing cues. In response, researchers have explored the integration of additional modalities—such as depth, infrared (IR), and thermal imaging—to provide richer and more discriminative information for robust spoof detection. Multimodal approaches offer complementary features that improve generalization and resilience against unseen attacks by leveraging diverse sensory inputs. However, conventional multi-stream architectures often process each modality independently, resulting in limited inter-modal interaction and redundant extraction of shared features, such as eye orientation. This lack of effective fusion can restrict the potential performance improvements gained through multimodal integration. Furthermore, increasing the number of modalities can substantially raise the computational complexity and parameter count of the system. Therefore, a key challenge remains: to design a unified framework that can jointly exploit complementary features across modalities while maintaining computational efficiency.

To address these challenges, transformer-based architectures, particularly Vision Transformers (ViTs), have emerged as powerful alternatives. Unlike CNNs, ViTs utilize self-attention to model long-range dependencies and global contextual relationships across image patches, making them well-suited for capturing subtle spoof cues. Thus, it is worth exploring ViT based multi-modal FAS.

4.1 Methodology

This section presents the proposed MF^{2}ShrT framework for FAS based on a pre-trained vision transformer (ViT), as illustrated in Figure 4.1. MF^{2}ShrT is a bi-

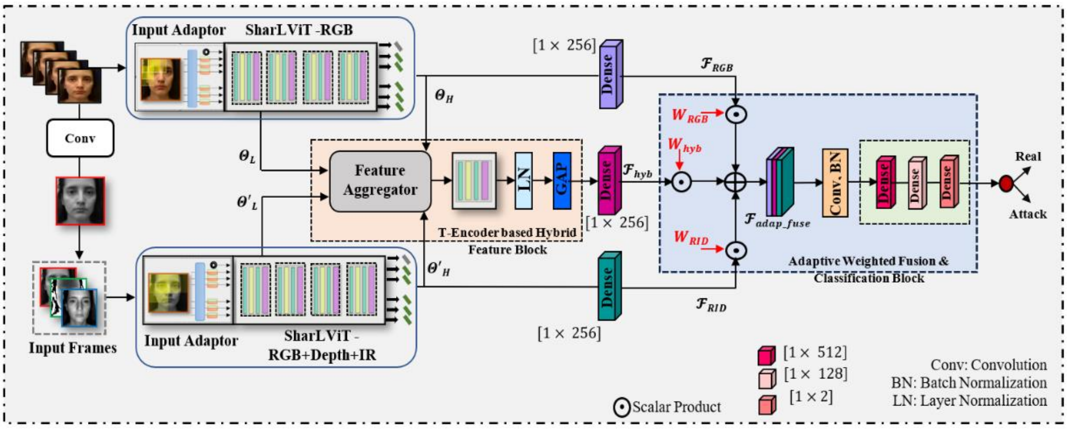


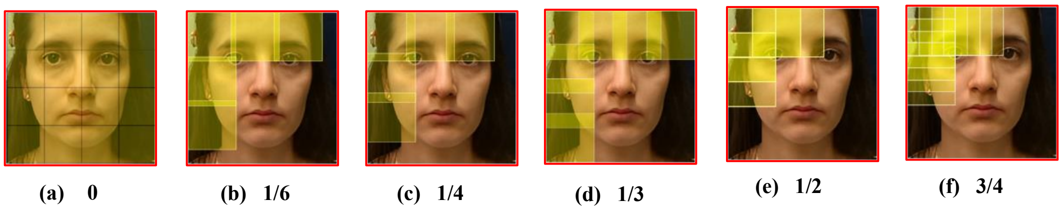
Figure 4.1 The Proposed framework MF^{2}ShrT for Face Anti-Spoofing

branch, bi-stage architecture that employs overlapped patches and parameter sharing to enhance efficiency. he first stage includes an input adaptor and a multi-modal feature extractor, while the second stage comprises a T-encoder-based hybrid feature block, adaptive weighted fusion, and a classification head.

One branch processes RGB images directly, while the second branch receives a composite input formed by projecting RGB into a single channel and stacking it with Depth and IR images—yielding a tri-channel input suitable for the ViT. This inputs design leverages complementary features across modalities without requiring separate transformers, reducing complexity while maintaining rich representation. Each transformer's output is passed through individual dense layers. Intermediate features from various ViT encoders are then aggregated and projected via the T-encoder block to capture cross-modal dependencies. These hybrid features are fused with branch outputs for final classification, resulting in a robust and efficient FAS system. The following sections describe each component of the framework in detail.

4.1.1 Input Adaptor

The standard ViT segments an input image into non-overlapping patches, which can overlook inter-patch dependencies and local contextual cues—crucial for FAS tasks. This rigid partitioning often leads to the loss of fine-grained facial features. To overcome this, we redefine the token generation process using sliding patches with overlapping regions, allowing better preservation of inter-patch relationships and local details.



(a) 0 (b) 1/6 (c) 1/4 (d) 1/3 (e) 1/2 (f) 3/4

Figure 4.2 Illustration of Overlapping Patch Rate for different values.

Specifically, given an input image $Z \in \mathbb{R}^{H\times H\times C}$, we extract overlapping patches using a patch size P and stride $S = P \times (1 - overlap\%)$. This results in a sequence of flattened $2D$ patches $Z_P \in \mathbb{R}^{N\times(P^2\times C)}$,where each image patch has resolution (P, P). The total number of patches is computed as:

$$N = [\frac{H+2\times\rho-(P-1)}{S} + 1] \tag{4.1}$$

where ρ denotes zero-padding and H represents input image resolution.

These flattened patches are projected to the transformer's model dimension D via a trainable linear projection, resulting in patch embeddings $Z_P E$. A learnable class token $Z_{CLS} = X_0^0$ is prepended to this sequence, and position embeddings are added to retain spatial information. The resulting input to the transformer encoder is:

$$X_0 = [Z_{CLS}; Z_P^1 E; Z_P^2 E; \ldots\ldots\ldots\ldots; Z_P^N E] + E_{Position} \tag{4.2}$$

This combined embedding X_0 is passed through L transformer encoder layers to generate the final face image representation X_L^0. Figure 4.2 visualizes different overlapping rates, and in our approach, we use a ½ overlap between consecutive patches.

4.1.2 SharLViT: Parameter Sharing in ViT

ViTs are particularly effective for handling high-resolution images and capturing both local and global dependencies across modalities [187]. A typical ViT comprises multiple encoder layers, each containing Multi-Head Self-Attention (MSA) and Multilayer Perceptron (MLP) blocks. These layers iteratively refine patch and class token embeddings. The feature transformation across layers can be described as:

$$y_l = MSA\big(LN(X_{l-1})\big) + X_{l-1}, l \in= 1, \ldots, L, \tag{4.3}$$

$$X_l = MLP\big(LN(y_l)\big) + y_l, l = 1, \ldots, L, \tag{4.4}$$

$$X_{out} = LN(X_L) \tag{4.5}$$

Here $X_{l-1} \in R^{(L+1)\times C}$ is the output of $(l-1)_{th}$ encoder and X_l will serve as input to

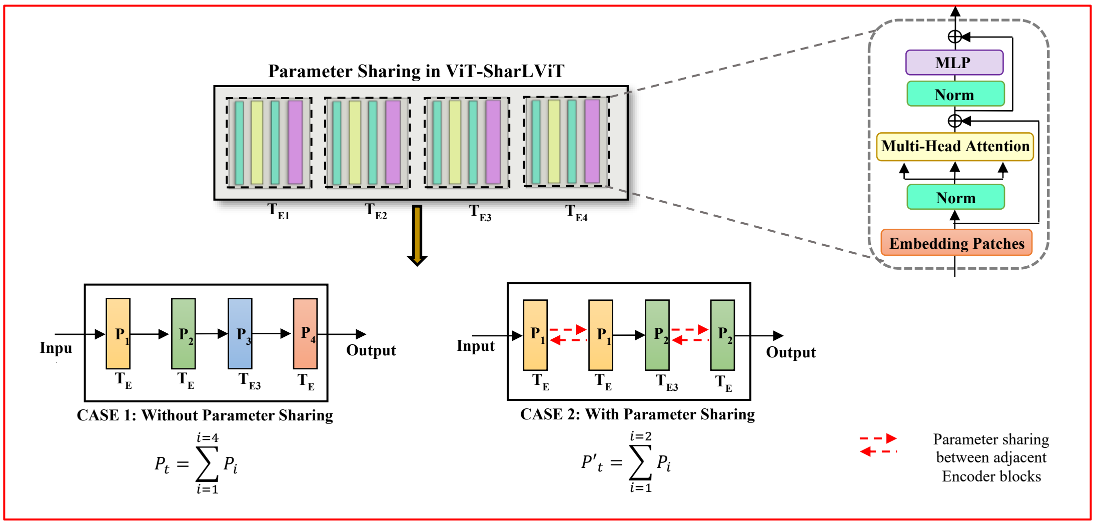

Parameter Sharing in ViT-SharLViT
T_E1
T_E2
T_E3
T_E4
MLP
Norm
Multi-Head Attention
Norm
Embedding Patches
Inpu
P_1
P_2
P_3
P_4
T_E
T_E
T_E3
T_E
Output
CASE 1: Without Parameter Sharing
$P_t = \sum_{i=1}^{i=4} P_i$
Input
P_1
P_1
P_2
P_2
T_E
T_E
T_E3
T_E
Output
CASE 2: With Parameter Sharing
$P'_t = \sum_{i=1}^{i=2} P_i$
Parameter sharing between adjacent Encoder blocks

Figure 4.3 SharLViT- Sequential Parameter Sharing in Vision Transformer Encoder Blocks. T_{E1}, $T_{E2,}$ T_{E3} and T_{E4} denote 4 encoder blocks within a Transformer. Case 1: Without Parameter Sharing- Each encoder block has parameters P_1, P_2, P_3 and P_4 respectively. Case 2: With Parameter Sharing: P_1 parameter is shared between encoders T_{E1} and T_{E2}, while the P_2 parameter is shared between encoders T_{E3} and $T_{E4.}$

the $(l+1)_{th}$ encoder and LN denotes the layer normalization operation. X_{out} is the final output embedding after normalization. This architecture enables ViT to capture subtle spoofing artifacts by modeling long-range and contextual interactions across facial regions.

In the proposed MF^{2}ShrT framework, one ViT processes RGB input, while another handles a composite tri-modal input (RGB, Depth, IR). This design allows MSA and MLP modules to jointly analyze cross-modal features, suppress redundant information, and enhance discriminative cues within each branch.

While increasing model size can improve performance [188], it also raises memory demands. To balance performance and efficiency, we adopt the ViT-B_16 [189] architecture, which contains 86M parameters across 12 layers—significantly lighter than larger counterparts like ViT-L_16 [189], ViT-L_32 [189], or ViT-H [189]. To further reduce computational overhead, we implement parameter sharing within the ViT backbone, resulting in the SharLViT architecture (see Figure 4.3).

Parameter sharing is applied sequentially: among N total transformer blocks, M blocks reuse a common set of parameters, reducing the number of distinct parameters sets to $J = \left[\frac{N}{M}\right]$. A representative transformation for any shared block is defined as:

$$y_i = MSA_j\big(LN(z_i)\big) + z_i, \quad i \in [1, M] \tag{4.6}$$

$$z_{i+1} = MLP_j\big(LN(y_i)\big) + y_i, \;\; j \in [1, J] \tag{4.7}$$

Here, MSA_j and MLP_j use shared parameters across blocks.

As shown in Figure 4.3, Case 1 illustrates a standard transformer with unique parameters P_i for each encoder block, resulting in a total parameter count $P_t = \sum_{i=1}^{i=n} P_i$.

In Case 2, parameter sharing reduces this to $P_t = \sum_{i=1}^{i=n/2} P_i$, effectively halving the parameter count.

MF^{2}ShrT employs four encoder blocks per branch $T_{E1:4}$ with M=2 and N=4, enabling efficient learning across eight total blocks using only four unique parameter sets. As shown in Section IV, our ablation study confirms that this parameter-sharing strategy maintains strong performance while significantly reducing computational cost.

4.1.3 T-Encoder-based Hybrid Feature Block

Traditional approaches [190] [135] often rely on deep layer stacking to capture semantic features; however, deeper networks tend to suppress low-level details. Capturing cross-level relationships between multi-modal features can improve performance. To this end, we propose a novel T-encoder-based Hybrid Feature Block that enhances global contextual understanding and bridges modality gaps commonly seen in CNN-based architectures [134] [136] [190].

Feature Aggregator: This block begins by extracting intermediate features $\mathcal{F}_{intermediate}$ from both RGB and RID branches using outputs from encoders T_{E2} and T_{E4} (Figure 4.4). The RGB branch outputs $\Theta_L \; and \; \Theta_H$ are concatenated, passed through convolution and global average pooling (GAP). In contrast, the RID branch output $\Theta'_L \; and \; \Theta'_H$ undergo individual convolution and GAP, preserving their discriminative strength from independent modalities. These are combined to form the intermediate feature map:

$$\mathcal{F}_{intermediate} = conc(\Theta_{RGB}, \Theta_{RID}^{L}, \Theta_{RID}^{H}) \tag{4.8}$$

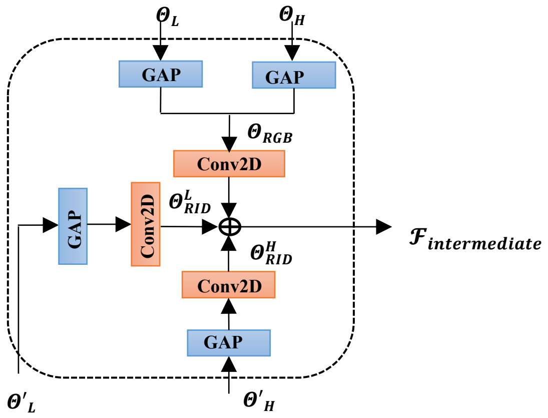


Figure 4.4 Detailed structure of feature aggregator**.**

where,

$$\Theta_{RGB} = conv(conc(GAP(\Theta_L), GAP(\Theta_H))) \qquad (4.9)$$

$$\Theta^L_{RID} = conv(GAP(\Theta'_L)) \qquad (4.10)$$

$$\Theta^H_{RID} = conv(GAP(\Theta'_H)) \qquad (4.11)$$

T-Encoder Feature Projector: It maps the aggregated feature representation, $\mathcal{F}_{intermediate}$, into a hybrid feature domain. This is done by reshaping $\mathcal{F}_{intermediate}$ and passing it through a uni-encoder transformer T_{E5}, followed by Layer Normalization (LN) and Global Average Pooling (GAP). This module is crucial for refining multimodal representations and identifying token importance by modeling complex intra- and inter-modal dependencies. The hierarchical structure of T_{E5} enables the extraction of high-level semantic and fine-grained cross-modal features from both branches. The output is then processed through a fully connected layer D to produce the hybrid feature map $\mathcal{F}_{hyb}$ of size $[1 \times\ 256]$, facilitating adaptive fusion with RGB and RID features. The overall operation is defined as:

$$\mathcal{F}_{hyb} = D\left(GAP\left(LN\left(T_{E5}\left(Conv(\mathcal{F}_{intermediate})\right)\right)\right)\right) \qquad (4.12)$$

4.1.4 Adaptive Weighted Fusion and Classification block

As described earlier, three distinct feature representations—$\mathcal{F}_{RGB}$, $\mathcal{F}_{hyb}$ and $\mathcal{F}_{RID}$ —are extracted from the input. Simple concatenation of these features [137] limits the network's ability to fully leverage their complementary strengths: w_{RGB}, w_{hyb} and w_{RID}. respectively. The adaptively fused representation $\mathcal{F}_{adap\_fuse}$ is formulated as:

$$\mathcal{F}_{adap\_fuse} = \left(conc\left((w_{RGB} * \mathcal{F}_{RGB}), (w_{hyb} * \mathcal{F}_{hyb}), (w_{RID} * \mathcal{F}_{RID})\right)\right) \quad (4.13)$$

Here, *conc* denotes the concatenation operation. The fused features are then passed through a classification pipeline comprising a convolutional layer, Batch Normalization (BN), and three dense layers with dimensions $[1 \times 512]$, $[1 \times 128]$ and $[1 \times 2]$. ReLU activation is applied to the first two dense layers to model non-linearities, while the final layer uses Sigmoid Focal Cross-Entropy to estimate the probability of the input being real or fake.

4.2 Database and Experimental Analysis

This section outlines the datasets, evaluation metrics, and implementation details. It also presents the performance of the proposed method, includes comparative analyses, and validates the framework through ablation studies.

4.2.1 Datasets

Experiments are conducted on two challenging datasets: CASIA-SURF (CS) [187] and WMCA [191]. The CS dataset includes 21,000 videos from 1,000 subjects, each with one live and six spoof video clips captured using Intel RealSense SR300 across RGB, depth, and IR modalities. Attack samples are created using A4 printouts with cutouts of facial regions in six variations. After sampling every tenth frame, the dataset is split into training (148K frames), validation (48K), and testing (295K) across 300, 100, and 600 subjects, respectively. Live and spoof samples 4,5,6 is used for training/validation, and 1,2,3 for testing. The WMCA dataset includes diverse 2D and 3D presentation attacks across four modalities, with two protocols: *seen* (known attacks) and *unseen* (generalization to new attacks).

4.2.2 Evaluation Metrics

The standard metrics used are APCER, BPCER, and ACER, defined as:

$$APCER = \frac{FP}{FP+TN} \quad (4.14)$$

$$BPCER = \frac{FN}{FN+TP} \quad (4.15)$$

$$ACER = \frac{APCER+BPCER}{2} \quad (4.16)$$

where FP, TN, FN, and TP denote false positives, true negatives, false negatives, and true positives. For cross-database testing, Half Total Error Rate (HTER) and the Area

Table 4.1 Performance on the WMCA dataset during Intra-dataset Testing (%).

Methods	APCER	BPCER	ACER
DeepPixBis [89]	8.2	3.7	6.0
MA-Net [141]	11.1	2.6	6.8
ResNet [184]	3.5	1.6	2.6
LBP-SVM [181]	8.5	0.6	4.6
MLP-Mixer [192]	1.7	2.3	2.0
Conv-MLP [145]	0.8	1.0	0.9
MF^{2}ShrT without T-Encoder branch (ours)	2.3	2.1	2.2
MF^{2}ShrT with T-Encoder branch (ours)	**0.11**	**0.13**	**0.12**

Table 4.2 Performance on the CASIA-SURF dataset during Intra-dataset Testing (%)

Methods	APCER	BPCER	ACER
Halfway Fusion [136]	5.6	3.8	4.7
SE Fusion [136]	3.8	1.0	2.4
Zhang et al. [187]	2.8	0.3	1.5
MA-Net [141]	2.4	1.7	2.0
Conv-MLP [145]	1.5	1.8	1.6
ViT+AMA+M^{2}A^{2}E [138]	0.81	0.42	0.62
MF^{2}ShrT without T-Encoder branch (ours)	4.1	2.9	3.5
MF^{2}ShrT with T-Encoder branch (ours)	**1.6**	**1.2**	**1.4**

Under the ROC Curve (AUC) are also reported.

4.2.3 Implementation Details

The framework employs the ViT-B_16 [189] as its backbone, initialized with ImageNet pre-trained weights. Custom dense layers are added using HeUniform initialization and ReLU [176] activation. Input image sizes are set to $[224 \times 224]$ or SharLViT-RGB and $[128 \times 128]$ for the SharLViT-RGB+Depth+IR variant. Experiments are conducted using Keras on a single NVIDIA P100 GPU (16GB RAM) via Google Colab Pro. Data augmentation includes rotation, random flips, and shuffling. The model is trained for 30 epochs with a batch size of 32, using the ADAM optimizer (learning rate: 0.0001) and Sigmoid Focal Cross-Entropy loss. The best-performing model is selected based on the lowest validation loss.

4.2.4 Comparative Analysis with other-state-of-the-arts

To assess the effectiveness of the proposed MF^{2}ShrT framework, we compare its performance with leading state-of-the-art methods. Tables 4.1–4.3 present intra-

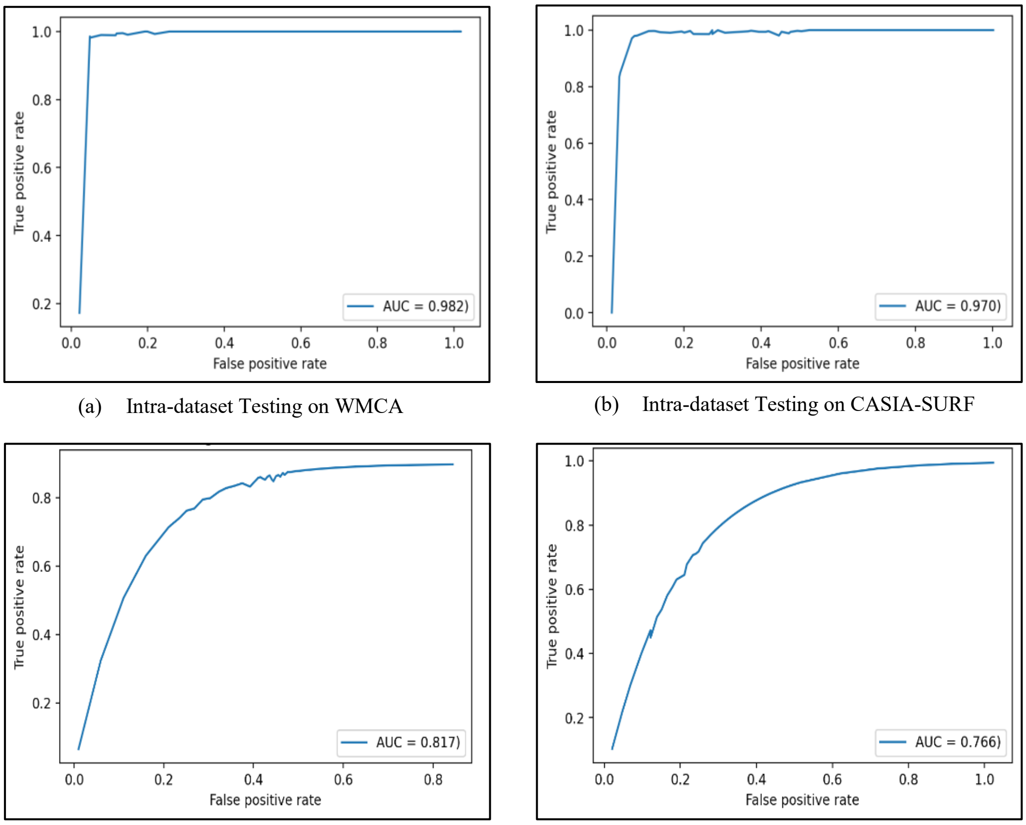
Table 4.3: Performance Evaluation for Cross-dataset Testing in terms of HTER (%).

Methods	**Train: CASIA-SURF Test: WMCA**	**Train: WMCA Test: CASIA-SURF**
Aux. (Depth) [93]	24.54	12.35
MM-CDCN [137]	21.83	21.25
MA-ViT [141]	20.63	10.41
ViT [189]	23.21	19.19
ViT+AMA+M^{2}A^{2}E [138]	18.83	8.60
MF^{2}ShrT without T-Encoder branch (ours)	20.21	14.13
MF^{2}ShrT with T-Encoder branch (ours)	19.43	13.84

and cross-dataset evaluation results on the WMCA and CASIA-SURF datasets, with the best scores highlighted in bold.

Intra-testing Results. Table 4.1 shows the performance on the WMCA dataset, comparing MF^{2}ShrT with notable approaches [89] [141] [181]**.** Our framework achieves the best performance with an ACER of just 0.12%, demonstrating its strength in reducing error rates. While competing methods primarily utilize CNN-based backbones, MF^{2}ShrT distinguishes itself as one of the few ViT-based solutions in face anti-spoofing, leveraging transformer architectures to model complex, multimodal dependencies effectively. Further, Table 4.2 evaluates MF^{2}ShrT on the CASIA-SURF dataset, in comparison with prominent multimodal frameworks [136] [138] [141] [145] [187]. With an ACER of 1.4%, MF^{2}ShrT performs competitively, outperforming most baselines and closely trailing the ViT+AMA+M^{2}A^{2}E [138] method. Notably, the latter incorporates modality-asymmetric masked autoencoders and self-supervised learning, yet our method achieves comparable accuracy without such complex pretraining. These results affirm the capability of our ViT-based architecture in capturing discriminative cues across modalities.

Cross-Dataset Testing To evaluate the generalization ability of MF^{2}ShrT we conduct cross-dataset experiments using the HTER metric, following the protocol in [138], Table 4.3 summarizes the outcomes. When trained on CASIA-SURF and tested on WMCA, our framework attains an HTER of 19.43%, securing the second-best performance. Conversely, training on WMCA and testing on CASIA-SURF yields an HTER of 13.84%, placing it fourth among competing methods. These results are



(a) Intra-dataset Testing on WMCA

(b) Intra-dataset Testing on CASIA-SURF

(c) Cross-dataset Testing: Train-WMCA and Test- CASIA-SURF

(d) Cross-dataset Testing: Train- CASIA-SURF and Test-WMCA

Figure 4.5 ROC curves for both Intra-dataset and Cross-dataset Testing on WMCA and CASIA-SURF.

Table 4.4 Ablation study of the effect of overlapping rate on the performance of WMCA dataset (%).

Overlapping Rate	ACER	Accuracy
0	7.45	92.86
1/6	10.85	89.21
1/4	3.55	96.60
1/3	1.5	97.97
1/2	**1.3**	**98.48**
3/4	**1**	**98.97**

notable, considering the substantial domain gap between the datasets. Although cross-dataset testing is inherently more challenging due to variations in attack types and image quality, MF^{2}ShrT exhibits performance that rivals top intra-dataset results—underscoring its robustness and transferability. Figure 4.5 further visualizes these outcomes through ROC curves, reinforcing the consistency of MF^{2}ShrT across diverse testing conditions.

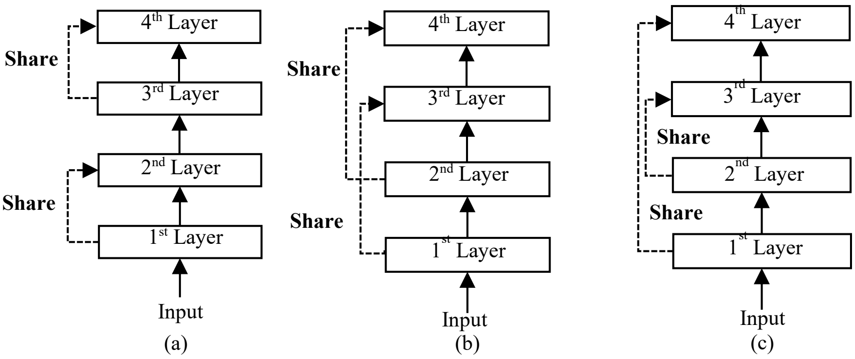


Figure 4.6 Types Parameter Sharing Strategies (a) Sequence Sharing (b) Cyclic Sharing (c) Reverse Cyclic Sharing.

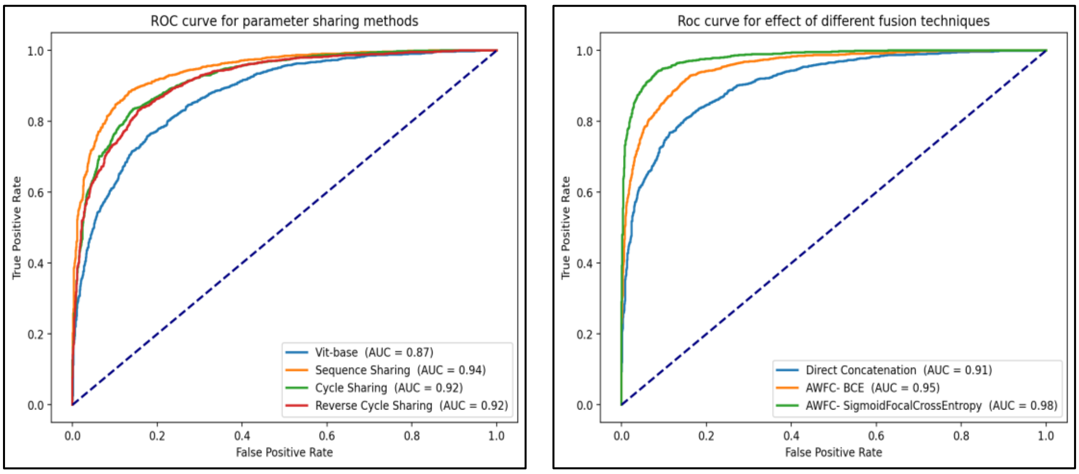
Table 4.5 Ablation study of baseline model and the effect of parameter sharing in the ViT model of WMCA dataset (%).

Model	APCER	BPCER	ACER
Vit-base	5.11	4.35	4.77
Sequence Sharing	**0.1**	**2.5**	**1.3**
Cycle Sharing	1.92	1.64	1.78
Reverse Cycle Sharing	3.19	2.51	2.88

4.2.5 Ablation Study

This section presents a comprehensive ablation study conducted on the WMCA dataset to evaluate the individual contributions of each key component in the proposed approach. All experiments are assessed using the ACER (%) metric.

Analysis of Overlapping Rate: Table 4.4 highlights the effect of varying overlap rates on classification performance using the WMCA dataset. Intuitively, increasing the overlap between patches tends to enhance performance. When no overlap is applied, the framework records a relatively high ACER, and performance declines further at a 1/6 overlap rate. However, beyond this point, the ACER begins to decrease, reaching its lowest value of 1.13% at a 3/4 overlap. While higher overlap produces more patches and captures richer spatial information, it also significantly increases computational cost due to a larger number of tokens processed by the transformer. Additionally, excessive overlap can introduce redundancy, as overlapping patches often contain similar information. Considering the minimal performance difference between the 1/2 and 3/4 overlap settings, and to maintain a balance between accuracy and efficiency,



(a) **(b)**

Figure 4.7 ROC curve for (a) parameter sharing methods, (b) Effect of weighted fusion classification (WFC) block.

Table 4.6 Ablation study of effect of different Fusion techniques on the performance of WMCA dataset.

Fusion Technique	**Initial Weights**			**Final Weights**			**ACER (%)**
	w_{RGB}	w_{hyb}	w_{RID}	w_{RGB}	w_{hyb}	w_{RID}	
With Direct Concatenation	-	-	-	-	-	-	3.57
Adaptively weighted fusion & Classification Block- BCE	0.66	0.22	0.12	0.14	0.55	0.31	1.7
	0.12	0.66	0.22	0.16	0.51	0.33	2.1
	0.22	0.12	0.66	0.15	0.49	0.36	1.3
	0.33	0.33	0.33	0.11	0.58	0.31	1.9
Adaptively weighted fusion & Classification Block - SigmoidFocalCrossEntropy	0.66	0.22	0.12	0.14	0.5	0.36	0.6
	0.12	0.66	0.22	0.13	0.52	0.35	0.9
	0.22	0.12	0.66	0.15	0.55	0.3	0.15
	0.33	0.33	0.33	0.18	0.55	0.37	0.12

the final framework adopts a 1/2 overlap rate.

Effectiveness of parameter sharing in ViT model: Three parameter-sharing techniques—Sequence, Cycle, and Reverse Cycle—were evaluated for internal layers in the ViT model, as illustrated in Figure 4.6. In the sequence sharing approach, identical parameters were assigned to every $[N/M]$ sequential layers. In the Cycle sharing, M uniquely parameterized layers were stacked and repeated in the same order until N layers were formed. In the Reverse Cycle, the stacking of M layers followed the Cycle order for M × ([N/M] − 1) layers, after which the remaining M layers were stacked in reverse order. Through parameter sharing, the overall model size and

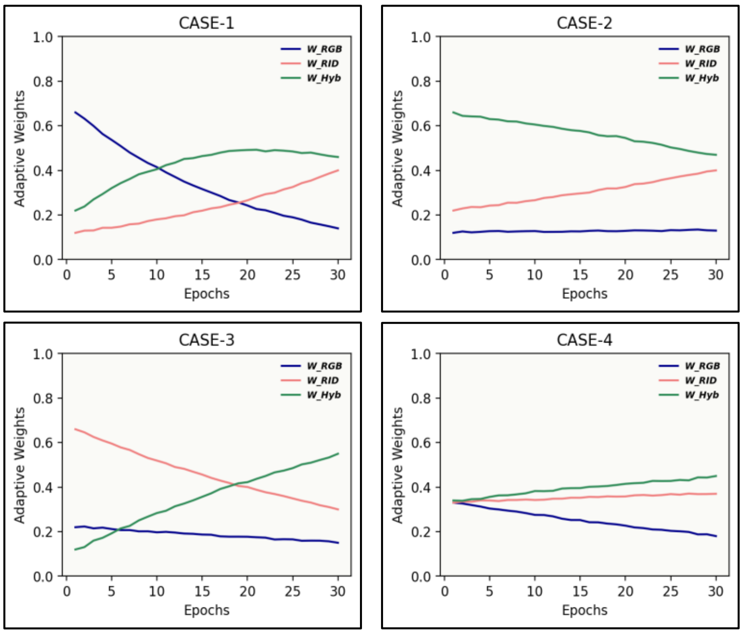


Figure 4.8 Curve of the adaptively assigned weighted over epochs

computational cost was reduced by a factor of [N/M]. Based on the experiments conducted, it was observed that Sequence sharing resulted in the lowest ACER compared to the baseline transformer and the other two sharing methods, as shown in Table 4.5. The ROC curve in Figure 4.7(a) further demonstrates the effectiveness of the Sequence sharing approach for face PAD, suggesting its usefulness in training scenarios involving large datasets.

Analysis of different types of Fusion techniques: Table 4.6 compares various fusion strategies, highlighting the effectiveness of the weighted fusion classification block. As shown in Figure 4.7(b), incorporating this block significantly improves performance. Since the contribution of the three branches varies, an adaptive fusion block was introduced to learn optimal weights during ablation testing. The final learned weights— $\{W_{RGB}, W_{hyb}, W_{RID}\}$ are $\{0.18, 0.51 \text{ and } 0.31\}$prioritize more discriminative features while minimizing the impact of redundant ones. Figure 4.8 illustrates the evolution of these adaptive weights over training epochs, demonstrating the performance of the weighted fusion and classification block using SigmoidFocalCrossEntropy across all four cases.

Table 4.7 Comparison the performance of the WMCA dataset using different losses (%).

Loss	APCER	BPCER	ACER
BCE	0.1	2.5	1.3
SigmoidFocalCrossEntropy	**0.11**	**0.13**	**0.12**

Table 4.8 Comparison of computation efficiency for various models on the WMCA dataset.

Method	Parameter	ACER (%)
ViTFAS [55]	85.8 M	7.56 ± 5.36
CDCN [70]	6.9 M	7.99 ± 5.51
MCCNN [191]	37.7 M	22.74 ± 15.33
ResNet [184]	42.5 M	19.56 ± 16.09
DenseNet [43]	26.4 M	21.02 ± 15.67
MLP-Mixer [192]	64.0 M	10.11 ± 12.05
FaceBagNet [135]	14.5 M	9.20 ± 9.99
Conv-MLP [145]	17.4 M	7.05 ± 11.16
MF^{2}ShrT (Ours)	37.9 M	6.81±14.67

Analysis of Losses: Two loss functions suitable for binary tasks were examined: Binary Cross Entropy (BCE) and Sigmoidal Focal Cross Entropy. The Focal Loss, introduced in the RetinaNet paper [193], is particularly effective for handling class imbalance.

The standard cross entropy (CE) loss is defined as:

$$CE(p, y) = \begin{cases} -log(p) & y = 1 \\ -lg(1-p) & otherwise \end{cases} \tag{4.17}$$

where $y \in [0,1]$ represents the ground truth and $p \in [0,1]$ is the model's predicted probability. To handle class imbalance, a class-weighted CE is used by introducing a weight parameter $\alpha \in [0,1]$ for class 1 and $1 - \alpha$ for class 0. The resulting class weighted CE can be expressed as:

$$CE(p, y) = \begin{cases} -\alpha * log(p) & y = 1 \\ -(1-\alpha) * log(1-p) & otherwise \end{cases} \tag{4.18}$$

However, α is fixed and not adaptable across different datasets. To overcome this, Focal Loss introduces a modulating factor:

$$FL(p, y) = \begin{cases} -(1-p)^{\theta} * log(p) & y = 1 \\ -(p)^{\theta} * log(1-p) & otherwise \end{cases} \tag{4.19}$$

Here, θ is a smoothing parameter that can be learned dynamically, allowing the model to adjust the loss based on sample difficulty during training. Our experiments show

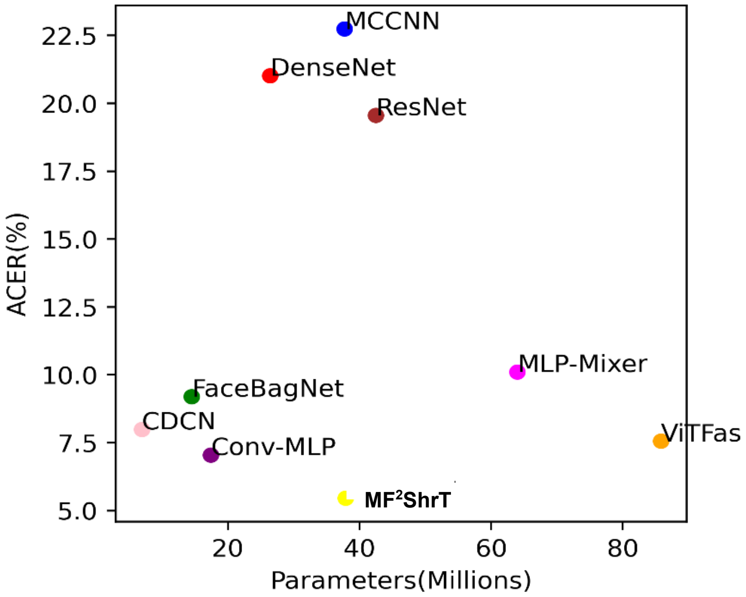


Figure 4.9 Trade-off between ACER and computational efficiency for various models under LOO protocol of WMCA dataset.

that using Sigmoidal Focal Cross Entropy improves performance, as reflected in Table 4.7.

Computational complexity: Based on the analysis presented in [145], the computational efficiency of the proposed framework was compared with existing state-of-the-art methods. As shown in Table 4.8, ACER values and trainable parameters were reported for each method. The base ViT model [55] and MLP-Mixer [192] were found to have the highest computational demands due to their large number of parameters. CDCN [70] exhibited significant computational complexity resulting from the calculation of feature map gradients. In [43] [184] [191], convolutional filters were applied across the entire image, imposing substantial computational load. In contrast, lower parameter count and reduced processing cost were achieved by FaceBagNet [135] using patch-based shared convolutions. Conv-MLP [145] incurred moderate computational requirements. All compared models, except ViTFAS, were CNN-based. The proposed framework was shown to require significantly fewer parameters than ViTFAS. Compared to CNN-based methods, it achieved the lowest ACER with a balanced computational cost.

In Figure 4.9, ACER versus computational efficiency is depicted using the

Leave-One-Out (LOO) protocol of the WMCA dataset. The proposed method, highlighted in yellow, was observed to offer competitive performance in both accuracy and efficiency. A throughput of ∼375 images per second (one per modality) was recorded, corresponding to an effective rate of 125 multimodal samples per second. This rate is 4–5 times higher than the typical framerate of consumer-grade FA cameras, indicating that the proposed framework can be effectively deployed in real-time applications.

4.3 Significant Outcomes

The outcomes of this chapter are summarized as follows:

- A multi-modal feature fusion framework, MF^{2}ShrT, is proposed, which employs overlapping patches and a shared-layer ViT to enhance local contextual information during feature extraction.
- A parameter-sharing mechanism, SharLViT, is introduced within the ViT architecture, effectively enhancing feature representation while significantly reducing model complexity and the number of trainable parameters.
- A novel T-Encoder-based Hybrid Feature Block is developed to capture inter-modal dependencies, enabling richer and more discriminative feature representations across modalities.
- Extensive evaluations conducted on the CASIA-SURF and WMCA datasets, using both intra- and cross-dataset protocols, demonstrate that the proposed framework achieves a strong balance between classification accuracy and computational efficiency when compared to state-of-the-art methods.
- Although the proposed framework maintains moderate computational cost, its modular and lightweight design makes it well-suited for real-time FAS applications.
- Despite achieving considerable success, existing methods—including the proposed framework—primarily focus on static features, often neglecting the temporal dimension, which plays a critical role in effective face anti-spoofing. Real and spoofed faces frequently appear visually identical in single-frame images, and ignoring temporal cues across consecutive frames leads to the loss

of essential motion-based information.

- The development of efficient and lightweight temporal networks remains an open research problem. The next chapter explores a temporal modeling framework to address this limitation by incorporating dynamic motion cues across video frames.

CHAPTER 5

BI-STAM: BI-DIRECTIONAL SPATIO-TEMPORAL ADAPTIVE MODELING FOR ROBUST FACE ANTI-SPOOFING

Recent studies leveraging convolutional neural networks (CNNs) [42] [73] have predominantly focused on extracting spatial features to classify live and spoofed faces. While these approaches have achieved notable success, they often overlook the temporal dimension, which plays a critical role in robust liveness detection. In many cases, real and spoofed faces appear visually similar in individual frames [194], and the absence of temporal modeling leads to the loss of vital motion cues [17] necessary for effective classification. To address this, temporal features such as eye-blinking, lip movements, and subtle facial motions have been explored to capture liveness signals across consecutive frames [167] [195]. However, early technique [42] [196] relying on hand-crafted temporal descriptors—such as Histogram of Oriented Optical Flow (HOOF) [167], Haralick features [197], or classical optical flow [198]—often lack the representational power needed to generalize across diverse and unseen presentation attacks (PAs). Deep learning-based temporal modeling has thus gained prominence, with methods employing recurrent neural networks (RNNs) [93] [199], 3D convolutional neural networks (3D CNNs) [200] [201], and two-stream architectures [37] [71] demonstrating improved capability in capturing temporal dependencies.

Despite these advancements, most existing solutions emphasize global facial dynamics while neglecting localized regions where subtle but crucial motion cues often reside. The unpredictable nature of these localized features [49][94][135] [202] —affected by pose, illumination, and material variations—necessitates a more comprehensive and adaptive strategy. Therefore, this chapter introduces a novel framework that explicitly models spatiotemporal relationships by leveraging both global and local motion information across frames. The objective is to effectively capture fine-grained temporal patterns embedded in localized facial patches, thereby enhancing the model’s ability to resist a wide range of sophisticated spoofing attempts.

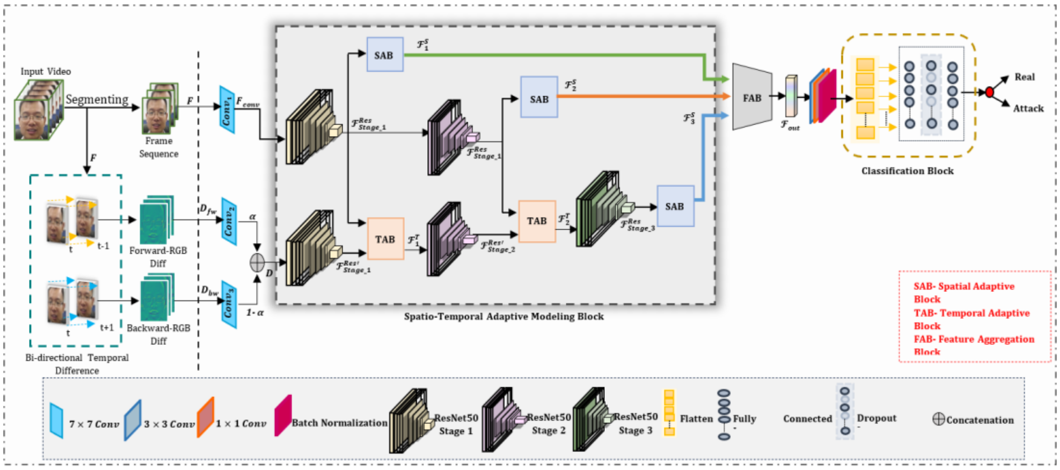


Figure 5.1 Illustration of the proposed Bi-STAM FAS framework

5.1 Methodology

The Bi-STAM framework is designed to robustly capture both spatial and temporal characteristics of facial motion for effective spoof detection. As illustrated in Figure 5.1, the framework is composed of the following key components: Bi-Directional Temporal Difference (BiD) Computation Spatio-Temporal Adaptive Modeling (STAM) Block Feature Aggregation Block (FAB) and Classifier**.** Each component is designed to work in synergy to model dynamic motion patterns and spatial saliency while keeping computational overhead low. The following sections describe each component of the framework in detail.

5.1.1 Bi-directional Temporal Difference (BiD)

A novel Spatio-Temporal Adaptive Modeling (STAM) block has been introduced within the Bi-STAM framework to enable robust FAS. Bi-directional temporal difference (BiD) computation is employed to extract intricate motion dynamics and spatially salient features from both forward and backward temporal directions. By utilizing dual-directional analysis, heightened sensitivity to subtle motion patterns is achieved, while computational efficiency is maintained. Dynamic facial motions are modeled with precision, and computational overhead is reduced in comparison to conventional approaches such as 3D CNNs or LSTM-based methods.

To process video-level information efficiently, a sparse sampling strategy [203] is employed. Each video is divided into *T* segments, and one frame is randomly selected from each segment. This sampling technique reduces the computational burden while preserving essential temporal information. Let the sampled frames be represented as F $\in [N, T, C, H, W]$ where N is the batch size, T is the number of segments, C is the number of feature channels, and H, W represent the frame dimensions. For each t^{th} frame its adjacent frames are denoted as $F_t^{(-1)}$ (preceding) and $F_t^{(+1)}$ (succeeding). The forward and backward temporal differences are computed as:

$$D_{fw} = \left\{F_t - F_t^{(-1)}\right\} \tag{5.1}$$

$$D_{bw} = \{F_t - F_t^{(+1)}\} \tag{5.2}$$

These bidirectional differences are then combined using convolutional operations and a learnable parameter α:

$$D = \alpha(Conv_2(D_{fw})) + (1 - \alpha)(Conv_3(D_{bw})) \tag{5.3}$$

Here $D \epsilon \mathbb{R}^{N \times T \times C \times H \times W}$, $Conv_2$ and $Conv_3$ denote convolution operations designed for static features. The parameter α is learned to balance the contributions of the forward and backward motion information.

5.1.2 Spatio-Temporal Adaptive Modeling Block (STAM)

Although BiD computations effectively capture motion cues, their integration with static appearance features is necessary for robust FAS. To achieve this, a Spatio-Temporal Adaptive Modeling (STAM) block is employed to process motion and appearance cues concurrently. Two distinct inputs—F_{Conv} and D—are provided to the STAM block and passed through the Temporal Adaptive Block (TAB) and Spatial Adaptive Block (SAB), as illustrated in Figure 5.2. These blocks collaboratively extract discriminative spatiotemporal features by leveraging the complementary nature of motion and appearance data.

Initially, $\mathrm{F_{Conv}}$ and D are processed using the first stage of ResNet-50 [184], producing feature maps $\mathcal{F}_{\mathrm{Stage\_1}}^{\mathrm{Res}}$ and $\mathcal{F}_{\mathrm{Stage\_1}}^{\mathrm{Res\prime}}$ respectively. These are then passed through the first TAB to capture dynamic and static facial cues, resulting in a temporal representation $\mathcal{F}_1^{\mathrm{T}}$. Simultaneously, the SAB operate on s $\mathcal{F}_{\mathrm{Stage\_1}}^{\mathrm{Res}}$, using channel- and spatial-wise attention to generate the spatial representation $\mathcal{F}_1^{\mathrm{s}}$. Both outputs are forwarded to the second stage of ResNet-50, producing updated features $\mathcal{F}_{\mathrm{Stage\_2}}^{\mathrm{Res}}$ and $\mathcal{F}_2^{\mathrm{S}}$ respectively. To enrich spatial encoding, an additional SAB is applied to $\mathcal{F}_2^{\mathrm{T}}$, producing high-level semantic features $\mathcal{F}_3^{\mathrm{s}}$. Through this hierarchical structure, both

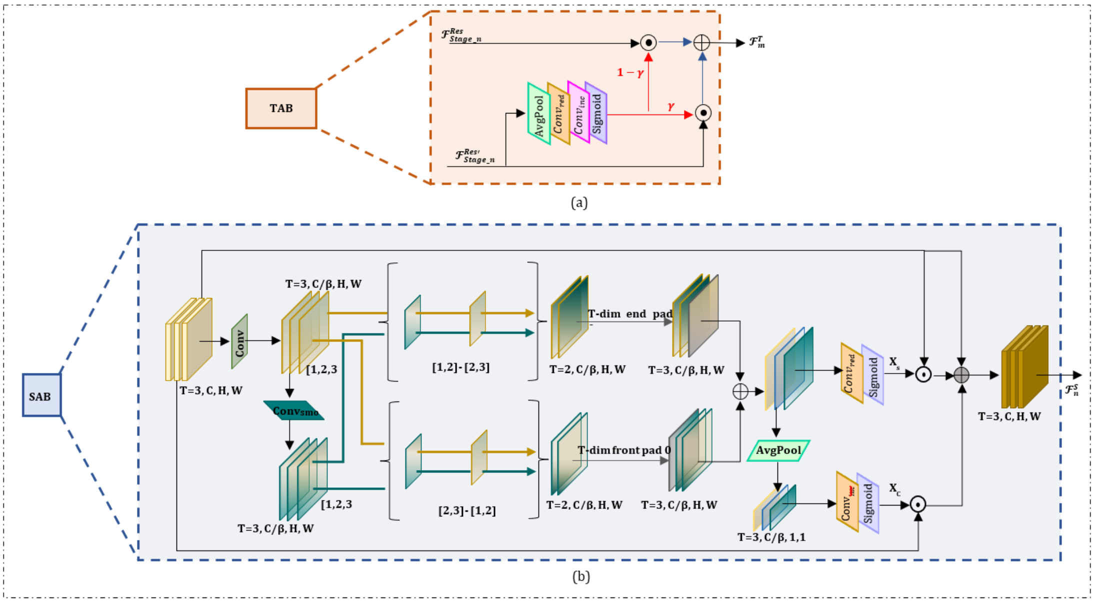


Figure 5.2 Illustration of (a) TAB: Temporal Adaptive Block, (b) SAB: Spatial Adaptive Block- An example with an input of three frames (T=3).

fine-grained and abstract spatiotemporal representations are captured, enhancing spoof detection performance.

Temporal Adaptive Modeling Block (TAB): The TAB, as illustrated in Figure 5.2(a), is designed to adaptively fuse motion dynamics with static semantic features. The TAB uses an adaptive fusion strategy that dynamically adjusts the fusion weight between motion dynamics and static semantics based on the input characteristics. It operates on inputs $\mathcal{F}^{Res}_{Stage\_n}$ and $\mathcal{F}^{Res\prime}_{Stage\_n}$, where n∈ {1,2} extracted from different stages of the ResNet-50 backbone. An adaptive fusion weight γ is computed by applying average pooling ($AvgPool$) to $\mathcal{F}^{Res\prime}_{Stage\_n}$, two sequential 1×1 convolutions ($Conv_{red}$ and $Conv_{inc}$) and a sigmoid activation:

$$\gamma = Sigmoid(Conv_{inc}\left(Conv_{red}\left(Avgpool(\mathcal{F}^{Res\prime}_{Stage\_n})\right)\right) \tag{5.4}$$

This adaptive weight modulates the fusion of static and dynamic features as $\mathcal{F}^{T}_{m}$:

$$\mathcal{F}^{T}_{m} = \left((1-\gamma)\odot\mathcal{F}^{Res}_{Stage_n}\right) + (\gamma\odot\mathcal{F}^{Res'}_{Stage_n}) \tag{5.5}$$

where $\odot$ denotes element-wise multiplication, and $n\,\epsilon$ [1,2,3] corresponds to the stages where TAB is applied. We have applied TAB in first two stages of the resNet-50 model resulting in temporal motion-enhanced representations $\mathcal{F}^{T}_{m}$ for $m\,\epsilon$ [1,2]. This fusion strategy enables the model to dynamically emphasize motion or appearance cues based on input characteristics, resulting in temporally enriched feature representations that enhance the model's ability to capture subtle facial motion patterns critical for spoof detection.

Spatial Adaptive Modeling Block (SAB): The Spatial Adaptive Block (SAB), shown in Figure 5.2(b), is designed to extract semantic features and enhance spoof-relevant spatial cues. Input features $\mathcal{F}_{Stage\_n}^{Res}$ are first downsampled via a 1×1 convolution by a factor of β, producing in the feature map $F \epsilon \mathbb{R}^{N\times T\times C/\beta\times H\times W}$. Temporal differences are computed using a 3 × 3 smoothing convolution ($Conv_{smo}$), generating forward and backward differences:

$$D_{fw} = \{Conv_{smo}(F_t) - F_{t-1}\}_{t=2}^{T} \tag{5.6}$$

$$D_{bw} = \{Conv_{smo}(F_t) - F_{t-1}\}_{t=1}^{T-1} \tag{5.7}$$

Zero-padding is applied to maintain the original shape, and the overall difference is computed as:

$$D = D_{fw} + D_{bw} \tag{5.8}$$

A spatial-wise attention map, X_s, is derived from D using a 1 × 1 convolution ($Conv_{red}$) followed by sigmoid activation and bias subtraction:

$$X_s = Sigmoid(Conv_{red}(D)) - \delta_s \tag{5.9}$$

where $X_s \epsilon \mathbb{R}^{N\times T\times 1\times H\times W}$ and δ_s is empirically set to 0.5.

Similarly, channel-wise attention, $\mathrm{X_c}$, is generated using average pooling and a 1 × 1 convolution ($\mathrm{Conv_{inc}}$):

$$X_c = Sigmoid(Conv_{inc}(AvgPool(D))) - \delta_c \tag{5.10}$$

where $X_c \epsilon \mathbb{R}^{N\times T\times C\times 1\times 1}$ and δ_c also set to 0.5.

These attention maps refine the spatial features via element-wise operations:

$$\mathcal{F}_n^s = \mathcal{F}_{Stage\_n}^{Res} + \mathcal{F}_{Stage\_n}^{Res} \odot X_c + \mathcal{F}_{Stage\_n}^{Res} \odot X_s \tag{5.11}$$

where $\mathcal{F}_n^s \epsilon \mathbb{R}^{N\times T\times C\times H\times W}$. The attention-enhanced features are integrated with the original features through element-wise multiplication and addition, further strengthening the model's ability to detect spoofing artifacts.

5.1.3 Feature Aggregator Block and Classification Block

Three feature representations, $\mathcal{F}_n^s$ ($n \epsilon$ [1,2,3]) were extracted at different

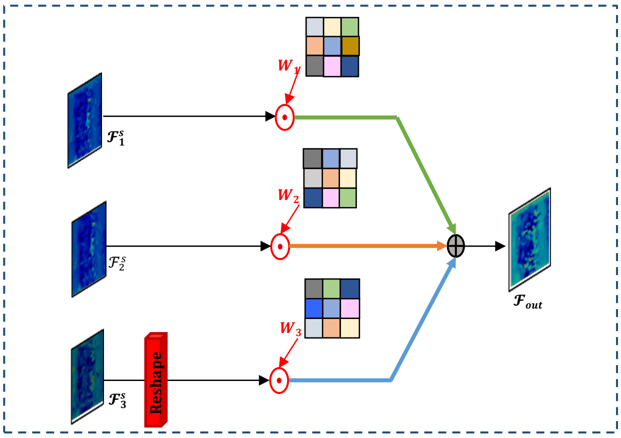


Figure 5.3 Illustration of Feature Aggregator Block (FAB**)**

network stages. To reduce redundancy and enhance feature learning, a Feature Aggregator Block (FAB) was introduced, as shown in Figure 5.3. Adaptive weights w_1, w_2, w_3 were assigned to the features $\mathcal{F}_1^s$, $\mathcal{F}_2^s$, and $\mathcal{F}_3^s$ respectively and the fused representation was computed as:

$$\mathcal{F}_{out} = \left(conc\left((w_1 * \mathcal{F}_1^S), (w_2 * \mathcal{F}_2^s), (w_3 * \mathcal{F}_3^s)\right)\right) \qquad (5.12)$$

where conc denotes concatenation. The fused features were refined using a 3×3 convolution for local spatial pattern learning, followed by a 1×1 convolution for channel reduction and consolidation. Batch normalization was applied to stabilize training and accelerate convergence. Finally, $\mathcal{F}_{\text{out}}$was passed through a classification block comprising flattening, a fully connected layer ($[1 \times 256]$) with ReLU activation, a dropout layer (0.5), and a final fully connected layer ($[1 \times 2]$). Binary Cross-Entropy loss was employed to compute the probability of the input being real or fake, enabling effective non-linear feature discrimination.

5.2 Database and Experimental Analysis

The preliminary work is initiated with a detailed description of the dataset, evaluation metrics, and implementation details of the proposed framework. Subsequently, evaluation results are presented along with comparative analyses. Finally, the framework's effectiveness is rigorously validated through ablation studies.

5.2.1 Datasets

The proposed model is evaluated using several widely adopted FAS benchmarks: OULU-NPU [204], MSU-MFSD [8], CASIA-MFSD [7], and Replay-

Attack [181].

The **OULU-NPU (O)** dataset is consists of recordings captured from six cameras across three sessions, featuring two printed and two replayed spoof types. Four protocols are defined: Protocols 1–3 evaluate performance across varying cameras, sessions, and spoof types, while Protocol 4 presents the most challenging scenario by assessing performance across all variations simultaneously.

The **MSU-MFSD (M)** dataset consists of 280 videos from 35 subjects, incorporating both photo and video-based attacks, including high-resolution and mobile phone replays, as well as printed photo attacks.

The **CASIA-MFSD (C)** dataset contains 600 videos with resolutions of 640×480 and 1280×720, recorded from 50 subjects using three different cameras. It is divided into training and testing sets of 20 and 30 subjects, respectively, and includes cut photo, warped photo, and video attacks.

The **Replay-Attack (RA)** dataset comprises 1200 videos from 50 subjects, with print, mobile, and high-definition attacks captured under both controlled and adverse lighting conditions.

5.2.2 Evaluation Metrics

To ensure a thorough and fair comparison with state-of-the-art methods, performance metrics specific to each dataset are utilized. For the CASIA-MFSD dataset, the framework is optimized on the training set and evaluated on the test set using the Equal Error Rate (EER), which balances false acceptance and rejection rates. For the Replay-Attack dataset, the Half Total Error Rate (HTER) is computed as the average of the False Rejection Rate (FRR) and False Acceptance Rate (FAR). On the OULU-NPU dataset, three primary metrics are employed: the Attack Presentation Classification Error Rate (APCER) for detecting spoof attempts, the Bona Fide Presentation Classification Error Rate (BPCER) for evaluating genuine attempts, and the Average Classification Error Rate (ACER), which is calculated as:

$$ACER = \frac{APCER+BPCER}{2} \tag{5.13}$$

For cross-database evaluations, HTER is employed to assess the model's generalization capability across different datasets. These metrics are used to provide a comprehensive framework for evaluating performance on various face anti-spoofing benchmarks, thereby enabling a detailed assessment of the framework's effectiveness. Additionally, the Area Under the Curve (AUC) is presented to further quantify overall performance.

5.2.3 Implementation Details

The Bi-STAM framework was developed using Keras and evaluated on Google Colab Pro equipped with NVIDIA P100 GPUs and 16 GB of RAM. It was built on a ResNet-50 [184] backbone, initialized with ImageNet pretrained weights from an open-source implementation. For optimal face detection and region of interest (ROI) extraction, the Viola-Jones algorithm[205]. was initially employed. Detected faces were standardized to [224 × 224 × 3] and used as RGB inputs. A sparse sampling strategy [203] was applied to select 16 or 32 frames from the original sequence in segments. The training was conducted over 60 epochs with a batch size of 8, using the ADAM optimizer and a learning rate of 0.0001 to minimize the Binary Cross-Entropy loss. Model selection was performed based on the lowest validation loss to ensure optimal performance.

5.2.4 Comparative Analysis with other-state-of-the-arts

To assess the performance of Bi-STAM, comparisons were conducted against several other methods. Results for both intra-dataset and cross-dataset testing are state-of-the-art (SOTA) methods. Results for both intra-dataset and cross-dataset testing are reported in Tables 5.1–5.5, with best-performing outcomes highlighted in bold.

Intra-Dataset Testing. Intra-dataset evaluations were carried out on CASIA-MFSD (C), Replay-Attack (RA), and OULU-NPU (O) datasets. Various baselines, including texture-based, CNN, GAN, and transformer-based models, were considered. Both frame-based [22] [73] [104] [151]and sequence-based methods [86], approaches were considered, with baselines including texture-based methods [22] [53] [151], CNNs [151] [206], transformers [86] [87], and GANs [104] [110]. On the CASIA-MFSD dataset shown in Table 5.1, superior performance was achieved by the proposed framework compared to leading SOTA methods. Similarly, Table 5.2 on the Replay-Attack dataset, the framework ranked first among the listed approaches. For OULU-

Table 5.1 Evaluation results on CASIA-MFSD (C) dataset (%).

Method	EER (%)
Color Texture [22]	6.20
Patch and Depth [206]	2.67
Attention [53]	3.14
FARCNN [151]	2.35
MIQF-SVM [207]	12.7
DTN [104]	1.34
DSCNN [45]	2.9
Zhang et al. [110]	1.17
Bi-STAM (ours)	**0.55**

Table 5.2 Evaluation results on Replay-Attack (RA) dataset (%).

Method	EER (%)	HTER (%)
Color Texture [22]	0.4	2.9
Patch and Depth [206]	0.79	0.72
Attention [53]	0.13	0.25
FARCNN [151]	0.06	0.18
MIQF-SVM [207]	-	5.38
DTN [104]	0.06	0.02
DSCNN [45]	4.7	0.39
Zhang et al. [110]	0.09	0.22
Bi-STAM (ours)	**0.02**	**0.13**

NPU, evaluations were conducted across all four protocols (Table 5.3). In Protocol 1, the proposed method outperformed six SOTA approaches; in Protocol 2, the third-best result was obtained. In Protocol 3, the method trailed behind TTN [86], Liu et. al. [208], and TransFAS [87] which utilize transformer and GAN-based designs. Nevertheless, the best ACER was reported in Protocol 4. These findings confirm the framework’s competitiveness within individual datasets. However, as intra-dataset testing is limited to specific conditions, cross- dataset analysis is further conducted to evaluate the robustness of the framework against unseen spoofing scenarios.

Cross-Dataset Testing. To evaluate the generalization capability of the proposed framework, cross-dataset testing was conducted using the CASIA-MFSD and Replay-Attack datasets. Two testing protocols were considered: training on CASIA-MFSD and testing on Replay-Attack (C→RA), and the reverse (RA→C). As reported in Table 5.4, the proposed Bi-STAM framework achieved the best performance in the RA→C protocol, outperforming existing state-of-the-art methods. In the C→RA scenario, it attained the second-best result, slightly behind DSCNN [45], which employs a multi-

Table 5.3 Evaluation results on Oulu-NPU(O) dataset (%).

Protocols	**Method**	**APCER (%)**	**BPCER (%)**	**ACER (%)**
1	DRL-FAS [94]	5.4	4.0	4.7
	CDCN [70]	0.4	1.7	1.0
	DAM [71]	1.4	1.8	1.6
	DTN [104]	0.78	1.06	0.92
	TTN [86]	1.2	0	0.6
	Liu et. al. [208]	0.6	0.0	0.3
	TransFAS [87]	0.8	0.0	0.4
	DSCNN [45]	0.37	2.9	1.6
	Zhang et al. [110]	0.63	0.80	0.72
	CSN-IR [209]	3.7	6.7	5.2
	Bi-STAM (ours)	**0.32**	**0.1**	**0.21**
2	DRL-FAS [94]	3.7	0.1	1.9
	CDCN [70]	1.5	1.4	1.5
	DAM [71]	2.6	0.8	1.7
	DTN [104]	3.84	2.11	2.88
	TTN [86]	0.8	0.8	0.8
	Liu et. al. [208]	0.7	1.4	1.1
	TransFAS [87]	1.5	0.5	1.0
	DSCNN [45]	3.1	7.2	5.2
	Zhang et al. [110]	2.53	1.36	1.95
	CSN-IR [209]	0	3.6	1.8
	Bi-STAM (ours)	**0.88**	**0.89**	**0.89**
3	DRL-FAS [94]	4.6±3.6	1.3±1.8	3.0±1.5
	CDCN [70]	2.4±1.3	2.2±2.0	2.3±1.4
	DAM [71]	2.0±2.6	3.9±2.2	2.8±2.4
	DTN [104]	1.9±1.6	3.8±6.4	2.8±2.7
	TTN [86]	0.8±0.9	1.4±1.8	1.1±0.9
	Liu et. al. [208]	1.5±1.3	1.4±1.3	1.5±1.1
	TransFAS [87]	0.6±0.7	1.1±2.5	0.9±1.1
	DSCNN [45]	5.6 ± 1.7	4 ± 3.3	4.8 ± 2.5
	Zhang et al. [110]	1.7±1.4	2.7±4.3	2.2±3.0
	CSN-IR [209]	8.6 ± 7.8	5.0 ± 8.3	6.8 ± 3.7
	Bi-STAM (ours)	**1.7±0.8**	**1.6±0.9**	**1.6±0.85**
4	DRL-FAS [94]	8.1±2.7	6.9±5.8	7.2±3.9
	CDCN [70]	4.6±4.6	9.2±8.0	6.9±2.9
	DAM [71]	4.2±5.2	4.6±3.8	4.4±4.5
	DTN [104]	4.0±4.1	3.0±4.9	3.5±2.4
	TTN [86]	4.2±2.4	3.8±4.0	4.0±2.3
	Liu et. al. [208]	4.2±3.0	1.7±2.6	3.0±1.9
	TransFAS [87]	2.1±2.2	3.8 ± 3.5	2.9±2.4
	DSCNN [45]	9.6 ± 6	7.8 ± 5.6	9.3 ± 6.3
	Zhang et al. [110]	2.1±4.5	5.7±4.9	3.9±3.2
	CSN-IR [209]	22.2±22.8	10.8±20.4	16.5±10.2
	Bi-STAM (ours)	**2.9±4.5**	**2.7±4.9**	**2.8±4.7**

scale inversion strategy and a two-stream network, resulting in higher computational complexity. To further assess generalization under more diverse conditions,

Table 5.4 Cross-Dataset Performance Evaluation: CASIA-FASD vs. Replay-Attack (HTER%).

Methods	**C→RA**	**RA→C**
STASN [49]	31.5	30.9
CDCN [70]	15.5	32.6
DAM [71]	27.4	28.1
DTN [104]	16.64	22.98
DRL-FAS [94]	28.4	33.2
Liu et. al. [208]	22.0	26.7
Zhang et al. [110]	25.73	21.57
DSCNN [45]	11.1	6.11
Bi-STAM (ours)	**14.5**	**4.35**

Table 5.5 Cross-Dataset Performance Evaluation: OULU-NPU(O), MSU-MFSD (M), CASIA-MFSD (C) and REPLAY-ATTACK (RA) in HTER (%) and AUC (%).

Method	**RA&C&M→ O**		**O&C&M→ RA**		**O&C&RA→ M**		**O&M&RA→ C**	
	HTER (%)	**AUC (%)**	**HTER (%)**	**AUC (%)**	**HTER (%)**	**AUC (%)**	**HTER (%)**	**AUC (%)**
Color Texture [22]	63.59	32.71	40.4	62.78	28.09	78.47	30.58	76.89
MADDG [210]	17.69	88.06	24.5	84.51	22.19	84.99	27.98	80.02
CDCN [70]	22.90	85.45	22.46	86.64	19.98	84.75	16.92	90.46
DRDG [211]	12.43	95.81	19.05	88.79	15.56	91.79	15.63	91.75
DR-UDA [113]	24.7	-	22.7	-	16.1	-	22.2	-
DTN [104]	18.26	89.40	21.43	88.81	19.40	86.87	22.03	87.71
TTN [86]	12.64	94.20	14.15	94.06	9.58	95.79	9.81	95.07
TransFAS [87]	7.08	96.69	9.81	96.13	10.12	95.53	15.52	91.10
Bi-STAM (ours)	**15.65**	**91.68**	**12.90**	**94.50**	**13.81**	**93.39**	**10.45**	**95.76**

Additional evaluations were carried out using three datasets for training and one remaining dataset for testing. These experiments encompassed four configurations: O&C&RA→M, O&M&RA→C, O&C&M→RA, and RA&C&M→O. The corresponding results, summarized in Table 5.5 using HTER and AUC metrics, demonstrate that Bi-STAM delivers strong performance across most settings. It ranked second in the O&C&M→RA and O&M&RA→C protocols, third in O&C&RA→M, and fourth in RA&C&M→O. The minor performance drop in certain settings is likely due to the higher temporal variability in the RA and OULU-NPU datasets, which facilitates temporal cue extraction, whereas the CASIA-MFSD and MSU-MFSD datasets exhibit limited motion and simpler recording conditions.

Table 5.6 Performance evaluation of various network configurations on Protocol 2 of the OULU-NPU dataset, using 8 selected frames (%).

TAB	SAB	BiD	Baseline	ACER (%)
-	-	-	ImageNet	12.02
√	-	√	ImageNet	11.86
-	√	√	ImageNet	11.35
√	√	-	ImageNet	9.2
√	√	√	ImageNet	**8.7**

Table 5.7 Comparative Performance Analysis of Various Connection Strategies for TAB and SAB (%).

Connection strategies	Configuration	ACER (%)
Cascaded	TAB (stage-1,2) + SAB (stage-3)	8.7
	SAB (stage-1,2) + TAB (stage-3)	10.53
Parallel	TAB (stage-1,2,3) + SAB (stage-1,2,3)	12.45
Combined	Cascaded {TAB (stage-1,2) + SAB (stage-3)} + Parallel SAB (stage-1)	**5.8**
	Cascaded {TAB (stage-1,2) + SAB (stage-3)} + Parallel SAB (stage-1,2)	**2.1**

These findings underscore the robust generalization capability of Bi-STAM across varying domains. Despite slight performance variations, Bi-STAM distinguishes itself from most static-input-based SOTA methods—except TTN [86] — by effectively leveraging temporal information. This temporal modeling enhances its resilience to a wide range of spoofing scenarios, positioning Bi-STAM as a meaningful advancement in face anti-spoofing research. Furthermore, the results highlight areas for future refinement, particularly in contexts involving complex motion patterns.

5.2.5 Ablation study

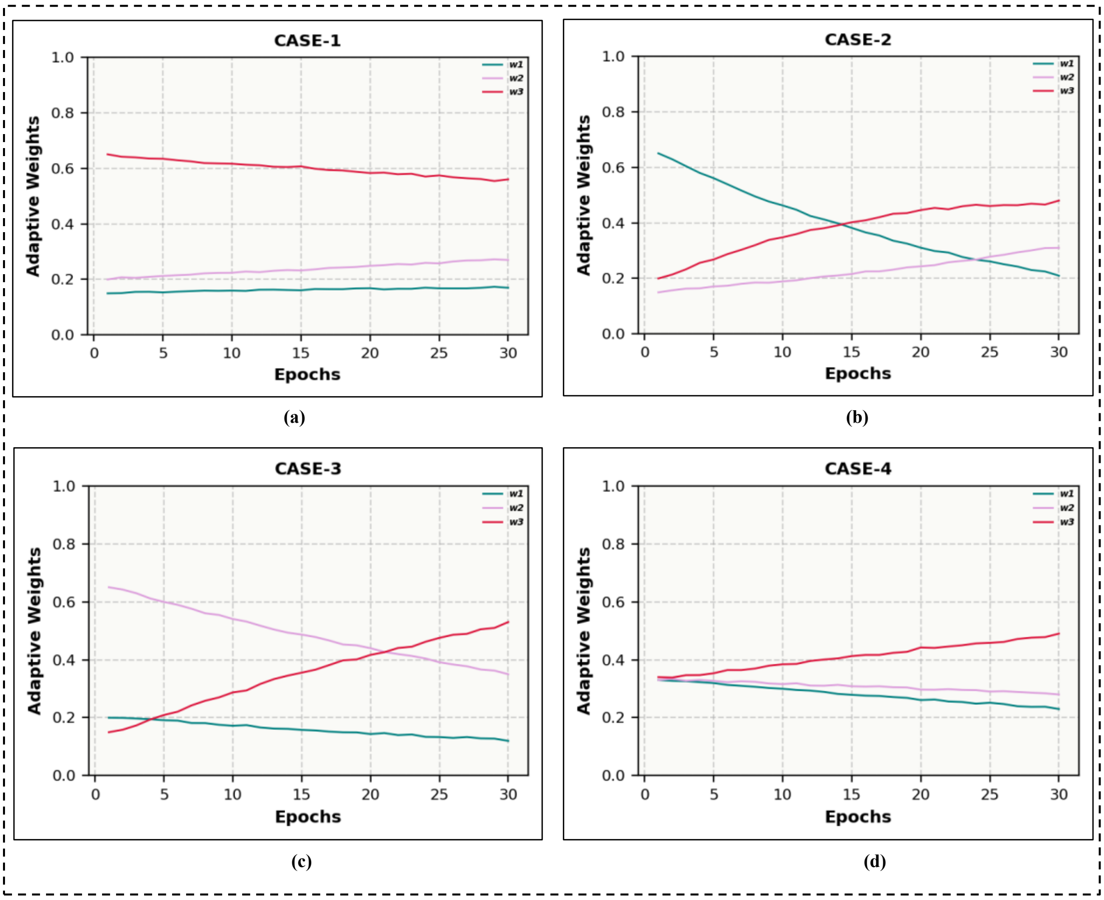
This section presents an ablation study to evaluate the individual contributions of each component in the Bi-STAM framework. The experiments are conducted on Protocol 2 of the OULU-NPU dataset, with performance measured using the ACER (%) metric.

Analysis of each component. The impact of each component in the Bi-STAM framework was evaluated using 8-frame video sequences. As detailed in Table 5.6 (where "√" denotes inclusion and "−" denotes exclusion), replacing the uni-directional temporal difference with a bi-directional approach improves ACER by over 1%, demonstrating enhanced motion representation. Adding the Temporal Attention Block (TAB) alone reduces ACER from 12.02% to 11.86%, effectively capturing low-level motion features. Integrating the Semantic Attention Block (SAB) lowers ACER to 11.35%, reflecting its strength in refining texture and semantic information. The combined use of TAB and SAB further improves ACER by over 2%, underscoring their complementary roles in motion and texture modeling.

Analysis of Connection Strategies for TAB and SAB. The effectiveness of TAB and SAB heavily depends on their placement within the network. Since shallow layers capture low-level motion and appearance cues, TAB is inserted at stages 1 and 2 of ResNet-50 to enhance motion modeling. In contrast, SAB is placed in stage 3 to refine high-level semantic features [212]. As shown in Table 5.7, positioning these modules in deeper layers degrades performance by suppressing crucial low-level information. The best ACER is achieved when TAB and SAB are cascaded up to stage 3, emphasizing the importance of strategic module placement for optimal temporal and semantic feature extraction.

Analysis of Multi-scale Spatial Features. To further enhance the framework, SAB was added in parallel to the cascaded configuration of TAB (stages 1–2) and SAB (stage 3). This setup reduced the ACER to 5.8%, as shown at the bottom of Table 5.7. Incorporating an additional SAB block for static feature extraction further improved performance, though adding more SAB blocks yielded no significant benefit. The outputs from all three SAB blocks were concatenated before the classification stage, highlighting the complementary roles of TAB and SAB in improving FAS effectiveness.

Analysis of Feature Aggregator block (FAB). Figure 5.4 illustrates the evolution of adaptive weights across epochs, highlighting their effect on Protocol 2 performance in the OULU-NPU dataset. Ablation studies revealed that the outputs from the three feature blocks contribute unequally. To address this, an adaptive fusion mechanism was introduced in the FAB to dynamically learn optimal weights. As shown in Figure5.5, the final weights assigned to features $\mathcal{F}_1^s$, $\mathcal{F}_2^s$, and $\mathcal{F}_3^s$ are $\{0.23, 0.28, 0.49\}$ respectively. This adaptive weighting prioritizes more discriminative features and



(a)

(b)

(c)

(d)

Figure 5.4 Adaptive weight progression over epochs: Initial weights $\{w_1, w_2, w_3\}$ are displayed, demonstrating their convergence to distinct values, indicating the respective contributions of $\mathcal{F}_1^s$, $\mathcal{F}_2^s$, and $\mathcal{F}_3^s$.

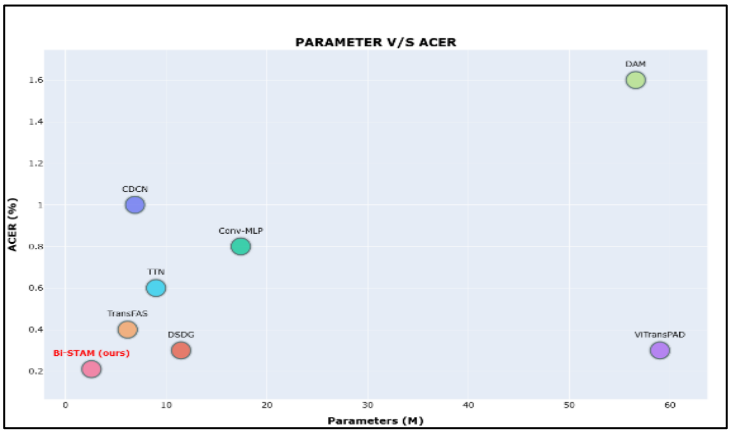


Figure 5.5 Visualization of ACER (%) on the OULU-NPU P1 dataset versus the number of parameters (M), highlighting the trade-off between model complexity and performance.

suppresses less informative ones, leading to more effective fusion and enhanced overall performance.

5.2.6 Analysis of Framework Complexity

The time and space complexity of an algorithm is crucial for its practical deployment. To assess this, we use parameter count (M) and Average Classification Error Rate (ACER %) as quantitative metrics of efficiency and performance. We compare the computational cost of our Bi-STAM framework against state-of-the-art methods. As shown in Figure 5.5 (OULU-NPU Protocol 1), Bi-STAM achieves an optimal trade-off, offering low ACER with fewer parameters. This efficiency underscores its ability to harness temporal information effectively, making it well-suited for real-world face anti-spoofing applications.

5.2.7 Visualization

Figure 5.6 illustrates Grad-CAM [213] visualizations for Bi-STAM, showing how attention maps evolve throughout training. Warmer colors indicate regions of higher attention, while cooler tones denote less focus. At epoch 5, the attention is broadly scattered, reflecting limited focus on discriminative facial regions. By epochs 15 and 55, attention becomes increasingly concentrated on key facial features—particularly in live images—while spoof images show more diffuse attention patterns. This progression indicates that the model initially struggles to identify relevant cues in spoofed data, but improves significantly as it learns to leverage motion cues, enhancing its spatio-temporal understanding for FAS.

To further illustrate the evolution of feature discrimination, we apply t-SNE [214] (Figure 5.7) on the training data from Protocol-1 of OULU-NPU, visualizing the outputs from the Feature Aggregation Block ($\mathcal{F}_{out}$) at various training stages. As training progresses, the separation between live and spoofed samples becomes more distinct, confirming that Bi-STAM increasingly emphasizes informative features while suppressing irrelevant ones, thus improving its ability to distinguish between genuine and spoofed faces.

5.3 Significant Outcomes

The outcomes of this chapter are summarized as follows:

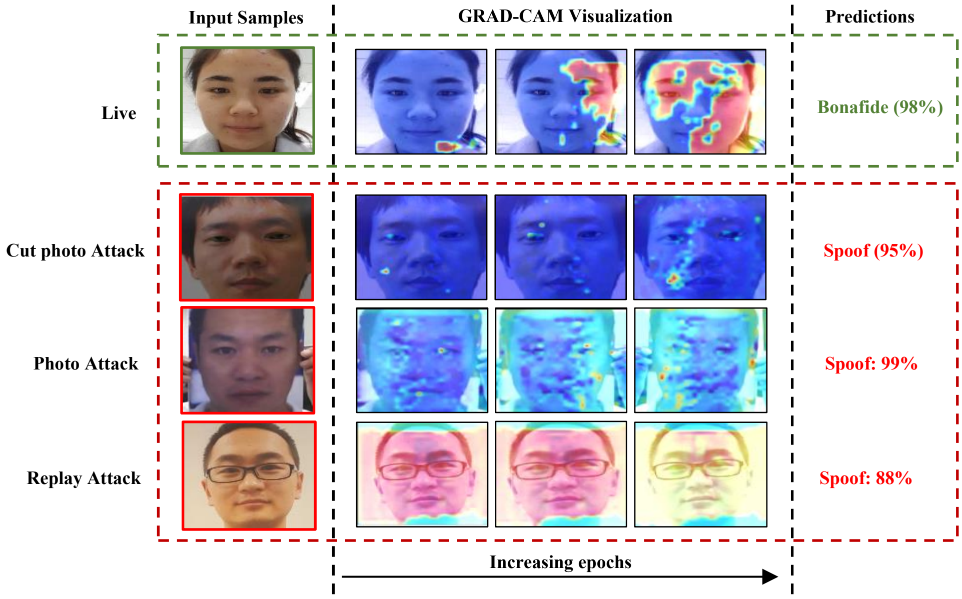
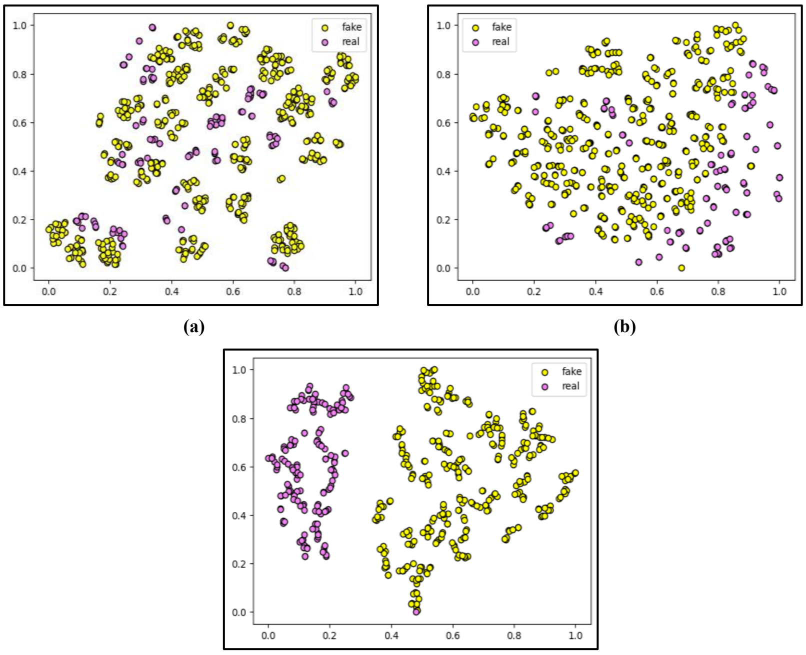


Figure 5.6 Visualization of Grad-CAM maps for the proposed Bi-STAM framework trained for 5, 15, and 55 epochs. Green-boxed samples represent live examples, while red-boxed samples indicate spoof examples. The first column presents the input samples, the second column shows the corresponding attention maps, and the third column displays the predictions for each sample.



(a)

(b)

(c)

Figure 5.7 t-SNE feature distributions for OULU-NPU Protocol 1 at various training stages: (a) Epoch 5, (b) Epoch 15, (c) Epoch 55.

- A bidirectional temporal difference (BiD) mechanism is introduced to capture motion cues from both forward and backward directions, enabling more effective modeling of temporal dynamics critical for face anti-spoofing (FAS).
- A novel Spatio-Temporal Adaptive Modeling (STAM) block is proposed, comprising a Temporal Adaptive Block (TAB) for learning dynamic motion patterns and a Spatial Adaptive Block (SAB) for refining spatial semantics through attention-enhanced feature extraction.
- An adaptive fusion mechanism within the Feature Aggregator Block (FAB) is developed to dynamically weight the outputs of TAB and SAB, improving feature selection by emphasizing discriminative information and suppressing redundancy.
- Extensive experiments are conducted on four benchmark datasets—OULU-NPU, CASIA-MFSD, MSU-MFSD, and Replay-Attack—demonstrating the framework's superior performance and generalization capabilities compared to state-of-the-art methods.
- While the proposed framework effectively models temporal and spatial cues, a critical limitation of existing methods—including Bi-STAM—is their reliance on global features, which often overlook subtle spoof patterns.
- To address this gap, the next chapter investigates fine-grained supervision using pixel-level annotations, aiming to capture localized spoof artifacts and further enhance spoof detection accuracy.

CHAPTER 6

GENERATIVE LEARNING-BASED PIXEL-WISE FACE ANTI-SPOOFING FRAMEWORK

In the previous chapters, we traced the progression of face anti-spoofing (FAS) techniques—from early handcrafted feature-based methods to deep learning-driven strategies, including static models, multi-modal architectures, and dynamic motion-aware approaches. While each paradigm significantly advanced spoof detection capabilities, persistent challenges remain in achieving robust generalization, resilience to unseen attacks, and fine-grained feature discrimination. These limitations have prompted a recent shift toward pixel-wise learning approaches, which offer a more granular and precise analysis of spoof-related cues. Pixel-wise methods operate at the per-pixel level, capturing detailed variations in depth, noise patterns, and residual features that are often imperceptible at coarser resolutions. This fine-level analysis enables a more accurate modeling of facial texture, geometry, and spoof artifacts. Among these, Generative Adversarial Network (GAN)-based models have shown substantial promise, particularly in learning discriminative representations and generating realistic depth maps through adversarial training, even without explicit spoof labels.

This chapter explores the foundations and recent advancements in pixel-wise FAS, with a particular focus on GAN-based frameworks that enhance both the granularity and generalization of spoof detection. The proposed work begins with **Case 1**, which utilizes RGB and MSCR (Multi-Scale Retinex with Color Restoration) as input to a GAN-based model, serving as the baseline. Building on this, **Case 2** refines the methodology by advancing MSCR to MSRCP (Multi-Scale Retinex with Color Preservation) to better preserve visual detail. Additionally, we introduce a novel Dual Polarized Self-Attention Guided Module (DPAttn) that adaptively evaluates and prioritizes RGB and MSRCP features. This guided fusion mechanism enhances the quality of generated depth maps, which are then leveraged by the generative network for improved live/spoof classification performance.

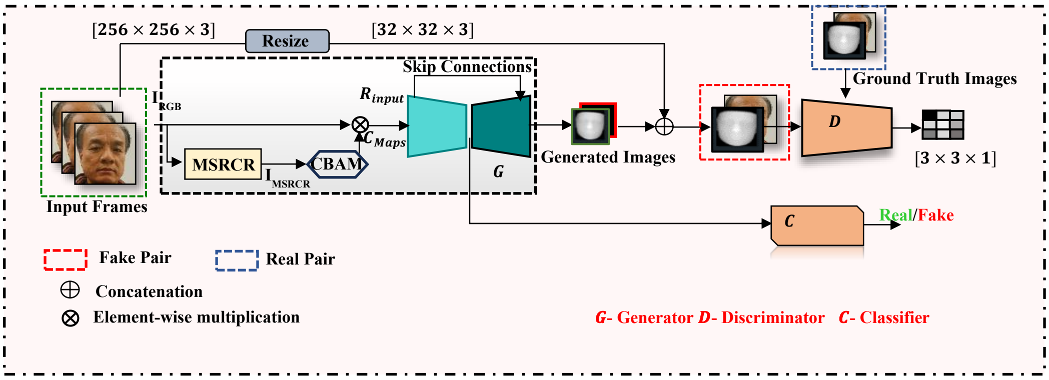


Figure 6.1 Overview of the proposed GAN-based framework

To evaluate the effectiveness and generalizability of the proposed framework, extensive experiments are conducted on four widely recognized face anti-spoofing benchmark datasets: MSU-MFSD [8], CASIA-FASD [7], REPLAY-ATTACK [181], and OULU-NPU [204].

To conduct a reliable comparison with prior research, we adhere to the original evaluation metrics associated with each benchmark dataset. For CASIA-FASD, model parameters are optimized on training sets and outcome is reported using the EER (Equal Error Rate) on the test set. In the case of Replay-Attack and ROSE-Youtu benchmark, we use the HTER (Half Total Error Rate). Within the OULU-NPU dataset, we follow ISO/IEC 30107-3 metrics [215] employing APCER (Attack Presentation Classification Error Rate), BPCER (Bona Fide Presentation Classification Error Rate), and ACER (Average Classification Error Rate). To ensure the generalizability, our principal evaluation criteria across all four datasets include the AUC (Area under curve) and the HTER (Half Total Error Rate).

CASE 1

6.1 Methodology

In the following section, we will provide a detailed explanation of the proposed framework as shown in Figure 6.1.

6.1.1 Multi-scale Retinex with Color Restoration (MSRCR)

To simulate the human visual system, particularly luminance perception, various image enhancement [216] algorithms have been explored. Land's Retinex theory [217] introduced a lightness model that laid the foundation for the Single Scale

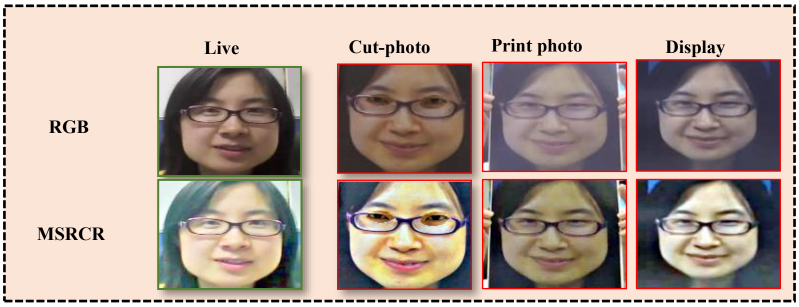


Figure 6.2 Pre-processes CASIA-FASD images: Green boundaries indicate live samples, red boundaries represent attack types. Top row shows RGB; Bottom row shows MSRCR.

Retinex (SSR) [218] which uses a Gaussian filter for illumination normalization. This was later refined by incorporating guided filtering [219].

To overcome SSR's scale limitations, the Multi-Scale Retinex (MSR) model [217] was proposed, combining outputs from multiple scales for improved enhancement. Adaptive-weight MSR [220] further normalized pixel values to the [0, 255] range for better contrast. However, MSR suffered from color distortion, which was addressed by Jobson et al. [217] through the integration of a Color Restoration (CR) function. This led to the development of MSRCR, which applies gain and offsets to each color channel, significantly improving color fidelity and contrast—especially under colored illumination.

MSRCR has demonstrated strong adaptability across datasets, enhancing visual cues critical for face anti-spoofing. Figure 6.2 presents RGB and MSRCR-enhanced images from the CASIA-FASD dataset, showcasing its effectiveness in improving feature visibility for both live and spoofed faces.

6.1.2 Convolutional Block Attention Module (CBAM)

The Convolutional Block Attention Module (CBAM) enhances feature learning by sequentially applying channel and spatial attention within the Generator framework, particularly on MSRCR-transformed RGB inputs denoted as $Ƒ\in\mathrm{R}^{\mathrm{C\times H\times W}}$. This procedure compresses the spatial dimension, yielding a 1D channel attention map $M_C\epsilon\mathbb{R}^{C\times1\times1}$. Subsequently, the modified tensor $Ƒ'$ traverses the spatial attention block, giving rise to a two-dimensional attention map $M_s\epsilon\mathbb{R}^{1\times H\times W}$. The refined output

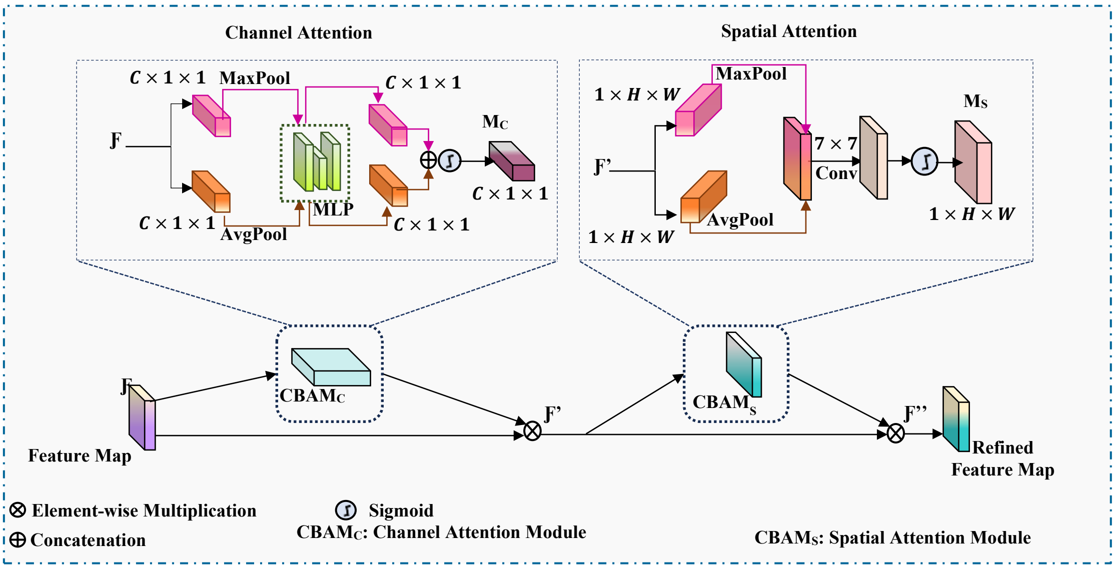


Figure 6.3 Schematic diagram of CBAM Module [3]

feature map is represented by the following equation:

$$\text{Ƒ}' = M_C(\text{Ƒ}) \otimes (\text{Ƒ}) \tag{6.1}$$

$$\text{Ƒ}'' = M_S(\text{Ƒ}') \otimes (\text{Ƒ}') \tag{6.2}$$

where $\otimes$ denotes the element wise multiplication. he architecture of CBAM, illustrated in Figure 6.3, comprises two key processes:

- ***Computation of Channel Attention***: This procedure refines fine-grained feature representations and mitigates information degradation by concurrently applying average and max pooling to capture compressed spatial dimensions. The derived descriptors, Ƒ^c_{avg} and Ƒ^c_{max} are passed through a shared MLP (Multilayer Perceptron) containing a single hidden layer. To minimize computational complexity and parameter usage, the hidden layer incorporates a reduction ratio of 8. The outputs of the MLP are then merged through element-wise addition, after which the sigmoid function is applied to generate the channel attention map $M_c(\text{Ƒ})\epsilon\mathbb{R}^{C\times1\times1}$, as specified in eq. (3):

$$\begin{aligned} M_C(\text{Ƒ}) &= \sigma\big(MLP(maxpool(\text{Ƒ})) + MLP(avgpool(\text{Ƒ}))\big) \\ &= \sigma(W_1(W_0(\text{Ƒ}^c_{max})) + W_1(W_0(\text{Ƒ}^c_{avg}))) \end{aligned} \tag{6.3}$$

  here, the symbol "+" symbolizes the element-wise addition.

- ***Computation of Spatial Attention:*** To emphasize informative regions in the spatial domain, this module performs average and max pooling along the channel axis, resulting in two 2D spatial descriptors: $\mathcal{F}^{s}_{max}\epsilon\mathbb{R}^{1\times H\times W}$and $\mathcal{F}^{s}_{avg}\epsilon\mathbb{R}^{1\times H\times W}$. These descriptors are concatenated and convolved using a 7x7 kernel, followed by a sigmoid activation to obtain the spatial attention map $M_s(\mathcal{F})\epsilon\mathbb{R}^{1\times H\times W}$:

$$\begin{aligned} M_s(\mathcal{F}) &= \sigma(f^{(7x7)}\{maxpool(\mathcal{F}); avgpool(\mathcal{F})\}) \\ &= \sigma(f^{(7x7)}\{\mathcal{F}^{s}_{max}; \mathcal{F}^{s}_{avg}\}) \end{aligned} \tag{6.4}$$

where, σ denotes sigmoid function, and f_{7x7}represents a convolutional operation with a 7x7 kernel. Equations (3) and (4) together describe the refinement process performed along both channel and spatial dimensions. This dual-attention mechanism significantly enriches the feature representations, thereby improving the overall feature extraction capability of the network.

6.1.3 Network Architecture

In Figure 6.3, the structure of the proposed GAN-based FAS framework is depicted. It is composed of three key components: a generator ($\boldsymbol{G}$) used for generating depth maps from RGB images, a discriminator ($\boldsymbol{D}$) tasked with evaluating the quality of the generator's outputs, and a classifier ($\boldsymbol{C}$) designed to distinguish between genuine and spoofed faces.

Generator. The generator, ($\boldsymbol{G}$), is implemented using an encoder-decoder structure with EfficientNetB4 [185] adopted as the base architecture to balance architecture capacity and computational efficiency**.** RGB images are encoded by the encoder, and the decoder is utilized exclusively for reconstructing depth maps. To prioritize multi-scale features, skip connections are integrated forming a U-Net architecture. Through these connections, representations from both the encoder and decoder are seamlessly merged facilitating smooth information flow. As a result, gradient propagation is improved and the vanishing gradient problem especially in the earlier layers is effectively mitigated.

Although deep learning methods possess strong non-linear feature learning capabilities, performance in anti-spoofing tasks is often degraded under varying input conditions. To overcome this issue, MSRCR images (I_{MSRCR}) are introduced by converting RGB images (I_{RGB}). In MSRCR, the luminance and chromatic components are processed independently, ensuring both illumination invariance and color preservation. Unlike the conventional RGB color space, which is sensitive to lighting variations, MSRCR isolates illumination information from color content. Consequently, a more robust representation against illumination changes is obtained. This representation is considered complementary to the RGB modality, which retains detailed yet illumination-sensitive features. To leverage the power of both domains, a CBAM mechanism is applied specifically on I_{MSRCR}, where essential features are emphasized based on their contextual and spatial relevance. The resulting attention map, C_{Maps}, assigns weights to each pixel in the image. An element-wise multiplication between C_{Maps} and I_{RGB}, is performed, producing the final refined input, R_{input}, as defined by:

$$R_{input} = I_{RGB} \otimes C_{maps} \tag{6.5}$$

This refined input is subsequently passed to the U-Net-based generator architecture, ensuring that informative features are extracted and propagated to the decoder blocks. A convolutional layer followed by a Tanh activation is employed at the end of the final decoder block to produce the output depth maps.

Discriminator. Within the framework, the discriminator ***D*** adopts a PatchGAN-inspired design [221], structured with a series of convolutional layers, each integrated with LeakyReLU activation and batch normalization. Two input pairs of dimensions $[32 \times 32]$, are processed: a real pair consisting of ground truth depth maps (D) with their corresponding RGB images, and a spoof pair consisting of generated depth maps $\{\text{Đ} = \mathrm{G(I)}\}$ with the same RGB images. This adversarial setup enables gradual improvement of G guided by the feedback provided through ***D*** gradients, resulting in more realistic depth map generation. The adversarial loss is defined as:

$$\mathbb{L}_{GAN}(\boldsymbol{G}, \boldsymbol{D}) = \mathbb{E}_{I,D}[log\boldsymbol{D}(I, D)] + \mathbb{E}_{I}[log(1 - \boldsymbol{D}(I, \text{Đ})] \tag{6.6}$$

To ensure stable training, additional image reconstruction losses such as the L1 norm are incorporated. The L1 reconstruction loss is expressed as:

$$\mathbb{L}_{L1}(\boldsymbol{G}) = \mathbb{E}_{I,D}[||D - \text{Đ}||_{1}] \tag{6.7}$$

The overall objective function for the proposed framework is formulated as:

$$\mathbb{L}_{our}(\boldsymbol{G}, \boldsymbol{D}) = arg_{\boldsymbol{G}}^{min\,max}{}_{\boldsymbol{D}}\mathbb{L}_{GAN}(\boldsymbol{G}, \boldsymbol{D}) + \lambda\mathbb{L}_{L1}(\boldsymbol{G}) \tag{6.8}$$

where λ serves as a regularization parameter to balance adversarial and reconstruction losses, enabling detailed and accurate generation of depth maps.

Classifier. The classifier constitutes the final component of the proposed FAS framework and is tasked with classifying input images as live or spoofed. It receives input from the latent representation extracted by the encoder of the generator, which, after GAN training, is expected to encode discriminative depth-aware features. Through this adversarial training, effective fusion of depth and RGB features is achieved by the encoder, thereby enhancing the generalization performance of the classifier. Optimization is carried out using the cross-entropy loss function, which guides the classifier in learning subtle patterns for accurate spoof detection. The loss is defined as:

$$\mathbb{L}_C = -(y\,log(p) + (1-y)\,log(1-p)) \tag{6.9}$$

where p represents the predicted probability and y denotes the ground truth label. This formulation encapsulates the classifier's objective of assessing the likelihood of an image being genuine or spoofed.

6.2 Experimental Analysis

In this section, a comprehensive overview of the datasets utilized and the evaluation metrics adopted for performance assessment is provided. The experimental setup is outlined, followed by the presentation of results, a comparative evaluation against recent benchmarks for each dataset, and a discussion of performance variations supported through an ablation study.

6.2.1 Implementation Details

Data Preprocessing. Frames are extracted at 10-frame intervals, with face detection initially using Viola-Jones, later replaced by MTCNN for higher accuracy. Ground-truth depth maps are obtained via PRNet— $[32 \times 32]$ for live faces and zeros for spoofed ones. Data augmentation includes random horizontal flipping to increase variability.

Training Setup. The proposed approach is developed in Keras and trained on Google Colab Pro equipped with an NVIDIA T4 GPU (16 GB RAM). The generator uses EfficientNetB4 [185] pretrained on ImageNet, and custom modules are initialized

Table 6.1 Intra-dataset assessment on the CASIA-FASD (C) Dataset (%).

Methods	EER
LBP [181]	18.2
Patch and Depth [73]	2.67
Attention [53]	3.14
Color Texture [22]	6.20
DTN [104]	1.34
ML-DAN [222]	3.7
FARCNN [151]	2.35
MIQF-SVM [207]	12.7
Zhang et al. [110]	**1.17**
DOG-ADTCP [223]	-
DSCNN [45]	2.9
Ours	**1.21**

Table 6.2 Intra-Dataset assessment on the Replay-Attack (RA) Dataset (%).

Methods	EER	HTER
LBP [181]	13.9	13.8
Patch and Depth [73]	0.79	0.72
Attention [53]	0.13	0.25
DTN [104]	0.06	0.02
ML-DAN [222]	0.3	0.6
MIQF-SVM [207]	-	5.38
Zhang et al. [110]	0.09	0.22
FARCNN [151]	0.06	0.18
DOG-ADTCP [223]	0.81	3.24
DSCNN [45]	4.7	0.39
Color Texture [22]	0.4	2.9
Ours	**0.05**	**0.03**

using HeUniform. Model optimization is performed using the Adam optimizer with an initial $1e-3$ learning rate, and a 16-batch size. The total loss function incorporates both adversarial and reconstruction losses, with the weighting parameters λ_{GAN} and λ_{L1} set to 1 and 100, respectively [104]. During each training epoch, image batches undergo random shuffling and horizontal flipping to further diversify the input distribution and improve generalization.

6.2.2 Comparative Analysis with other-state-of-the-arts

Intra-dataset Testing. We rigorously evaluated our proposed approach on three database—Replay-Attack (RA), OULU-NPU (O) and CASIA-FASD (C)—by adhering to their official protocols and comparing against other state-of-the-art methods [22] [53] [73] [104] [151] [181] [222]. Table 6.1 reports the EER for the C dataset, while Table 6.2 presents both EER and HTER values for the RA dataset. In both cases, our approach consistently outperforms prior methods, demonstrating its

Table 6.3 Intra-Dataset evaluation on OULU-NPU (O) (%).

Protocols	**Methods**	**APCER**	**BPCER**	**ACER**
1	Auxiliary [93]	1.6	1.6	1.6
	CPqD [224]	2.9	10.8	6.9
	CDCN [70]	0.4	1.7	1.1
	DSCNN [45]	0.37	2.9	1.6
	STASN [49]	1.2	2.5	1.9
	MIQF-SVM [207]	6.9	1.5	4.2
	SGTD [81]	2	**0**	1
	DTN [104]	0.78	1.06	0.92
	FaceDs [102]	1.2	1.7	1.5
	Zhang et al. [110]	0.63	0.80	0.72
	Ours	**0.3**	**0.9**	**0.6**
2	FaceDs [102]	4.2	4.4	4.3
	Auxiliary [93]	2.7	2.7	2.7
	CPqD [224]	14.7	3.6	9.2
	SGTD [81]	2.5	1.3	1.9
	STASN [49]	4.2	0.3	2.2
	CDCN [70]	1.5	1.4	1.5
	DSCNN [45]	3.1	7.2	5.2
	MIQF-SVM [207]	7.8	1.4	4.6
	DTN [104]	3.84	2.11	2.88
	Zhang et al. [110]	2.53	1.36	1.95
	Ours	**2.5**	**1.1**	**1.8**
3	Auxiliary [93]	2.7±1.3	3.1±1.7	2.9±1.5
	CPqD [224]	6.8±5.6	8.1±6.4	7.4±3.3
	FaceDs [102]	4.0±1.8	3.8±1.2	3.6±1.6
	CDCN [70]	2.4±1.3	2.2±2.0	2.3±1.4
	STASN [49]	4.7±3.9	0.9±1.2	2.8±1.6
	DSCNN [45]	5.6±1.7	4±3.3	4.8±2.5
	SGTD [81]	3.2±2.0	2.2±1.4	2.7±0.6
	DTN [104]	1.9±1.6	3.8±6.4	2.8±2.7
	MIQF-SVM [207]	3.6±0.9	4.3±1.8	4.0±1.4
	Zhang et al. [110]	1.7±1.4	2.7±4.3	2.2±3.0
	Ours	**1.6±1.1**	**2.5±1.0**	**2.05±1.1**
4	CPqD [224]	32.5±37.5	11.7±12.1	22.1±20.8
	FaceDs [102]	1.2±6.3	6.1±5.1	5.6±5.7
	STASN [49]	6.7±10.6	8.3±8.4	7.5±4.7
	CDCN [70]	4.6±4.6	9.2±8.0	6.9±2.9
	SGTD [81]	6.7±7.5	3.3±4.1	5.0±2.2
	Auxiliary [93]	9.3±5.6	10.4±6.0	9.5±6.0
	Zhang et al. [110]	2.1±4.5	5.7±4.9	3.9±3.2
	MIQF-SVM [207]	6.2±4.3	4.9±3.7	5.6±4.0
	DTN [104]	4.0±4.1	3.0±4.9	3.5±2.4
	DSCNN [45]	9.6±6.0	7.8±5.6	9.3±6.3
	Ours	**3.8±2.5**	**3.2±4.6**	**3.5±3.5**

robustness. For the O dataset, Table 6.3 summarizes results across all four protocols. Our method achieves a 0.6% low ACER of Protocol 1. In Protocol 2, the performance

Table 6.4 Comparative Cross-Dataset Analysis: CASIA-FASD vs. Replay-Attack (HTER %).

Methods	Train- C/Test- RA	Train- RA/Test-C
Deep-Learning [225]	48.2	45.4
LBP [181]	55.9	47.9
FARCNN [151]	26.0	29.4
LBP-TOP [226]	49.7	60.6
Auxiliary [93]	27.9	28.4
Color Texture [22]	47.0	39.6
STASN [49]	31.5	30.9
Zhang et al. [110]	25.73	21.57
DTN [104]	16.64	22.98
Attention [53]	30.0	33.4
Ours	**15.8**	**21.17**

Table 6.5 Comparative Cross-Dataset analysis across four datasets (%).

Test Case 1: C&RA&O→ M; **Test Case 2**: M&RA&O→ C; **Test Case 3:** C&M&O→ RA; **Test Case 4**: C&RA&M→ O

Methods	Test Case 1		Test Case 2		Test Case 3		Test Case 4	
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
MMD-AAE [227]	27.08	83.19	44.59	58.29	31.58	75.18	40.98	63.08
LBP-TOP [226]	36.9	70.80	42.6	61.05	49.45	49.54	53.15	44.09
DTN [104]	19.40	86.87	22.03	87.71	21.43	88.81	18.26	89.40
MADDG [210]	17.69	88.06	24.5	84.51	22.19	84.99	27.98	80.02
CDCN-PS [79]	20.42	87.43	18.25	86.76	19.55	86.38	15.76	92.43
Auxiliary [93]	22.72	85.88	33.52	73.15	29.14	71.69	30.17	77.61
Binary CNN [46]	29.25	82.87	34.88	71.94	34.47	65.88	29.61	77.54
CDCN [70]	22.90	85.45	22.46	86.64	19.98	84.75	16.92	90.46
Color Texture [22]	28.09	78.47	30.58	76.89	40.4	62.78	63.59	32.71
Ours	16.3	90.7	21.17	83.55	23.21	85.7	17.65	89.69

is on par with [70] only a marginal difference. Under Protocols 3 and 4, our framework attains the best results with ACER values of 2.05±1.1% and 3.5±3.5%, respectively. These outcomes underscore the competitiveness of our framework in both standard and challenging evaluation settings, and further emphasize the superior generalizability of our GAN-based depth map estimation compared to previous approaches.

Cross-dataset Testing. To evaluate the generalization ability of the proposed work across different domains, cross-dataset evaluation was conducted using the datasets CASIA-FASD (C) and Replay-Attack (RA). Two experimental protocols were

Table 6.6 Ablation Study Results: OULU-NPU Protocol 2 (%).

Backbone	Input	APCER	BPCER	ACER
Simple Encoder-Decoder (w/o skip connections)	RGB	11.5	8.9	10.2
U-Net	RGB	9.6	7.7	8.65
	RGB+MSRCR	5.4	4.8	5.1
With CBAM	RGB+MSRCR	2.5	1.1	1.8

considered: one with the C dataset used for training and RA for testing, and the other with the reverse arrangement. As shown in Table 6.4, our method achieves outstanding performance in both scenarios, confirming its strong cross-dataset robustness.

To comprehensively assess the generalization ability of the proposed work, cross-dataset evaluation was conducted across four diverse datasets, resulting in four distinct test cases, as summarized in Table 6.5. In each scenario, one dataset was assigned as the test set, and the remaining were utilized for training. The four test cases were defined as follows: Test Case 1 - O&C&RA → M, Test Case 2 - O&M&RA → C, Test Case 3 - O&C&M→RA, and Test Case 4 - RA&C&M→O. The outcomes reveal that our work achieves improved performance in Test Case 1, surpassing existing methods. In Test Case 2, it ranks second, following CDCN-PS [79] which leverages contrastive learning. In Test Case 4, our approach ranks third whereas in Test Case 3, it ranks fifth, performing slightly below CDCN [70], with only a marginal difference from both CDCN-PS [79] and CDCN [70]. Overall, these outcomes underscore the effectiveness of GAN-based framework in advancing the state of FAS.

6.2.3 Ablation Study

The ablation study was conducted on OULU-NPU Protocol 2, as shown in Table 6.6 to evaluate key components of the proposed framework. A basic encoder-decoder resulted in 10.2% of high ACER due to ineffective depth map generation. When replaced with a U-Net architecture, performance was improved, with an ACER of 8.65% using RGB input, and further reduced to 5.1% by incorporating MSRCR, highlighting the importance of input selection. However, challenges in complex video scenarios persisted. To address this, a CBAM attention module was integrated, leading to a significant improvement. The final model achieved an ACER of 1.8%, demonstrating that attention mechanisms play a crucial role in enhancing robustness and handling complex inputs.

6.2.4 Visualization and Analysis

Figure 6.4 illustrates representative visualizations produced by the proposed work for both real and spoof samples. Green boundaries denote live input samples, while red boundaries indicate fake samples. For genuine cases, the framework effectively reconstructs depth maps that align well with the ground truth, showing only slight deviations in fine details. In contrast, spoof samples primarily yield near-zero depth representations, sometimes appearing as noise-like patterns.

Figure 6.4 further illustrates certain failure instances. For example, although some inputs are genuine faces, the framework occasionally generates inaccurate depth maps. Similarly, in the case of a spoof face created using a replay video, the framework produces depth patches resembling actual facial structures rather than the expected zero-depth maps. Such cases reveal inherent challenges that can lead to classification errors. To investigate the distinctive ability of the CNN characteristics derived for FAS, here we employ the t-SNE visualization method [214]. By projecting the classifier's CNN features into a lower-dimensional space, clear separations between actual and spoof inputs can be observed, as shown in Figure 6.5. Specifically, Figure 6.5(a) depicts feature representations for the CASIA-FASD dataset with a framework trained on the same CASIA-FASD, whereas Figure 6.5(b) presents the distributions for the Replay-Attack using a framework trained on Replay-Attack. These results demonstrate that the proposed framework effectively transforms RGB face inputs into the depth domain, yielding consistent feature representation for real and spoof samples across diverse datasets.

These findings indicate that although the proposed work is capable of learning a broad set of features for depth map generation, it faces challenges in reliably distinguishing genuine faces from spoofed ones. This limitation can be attributed to the reliance on a standard U-Net generator, which may be insufficient for capturing the subtle and complex features necessary for accurate depth estimation across all cases. Since this represents an initial experimental effort, there remains considerable potential for improvement. Future research could focus on integrating more advanced

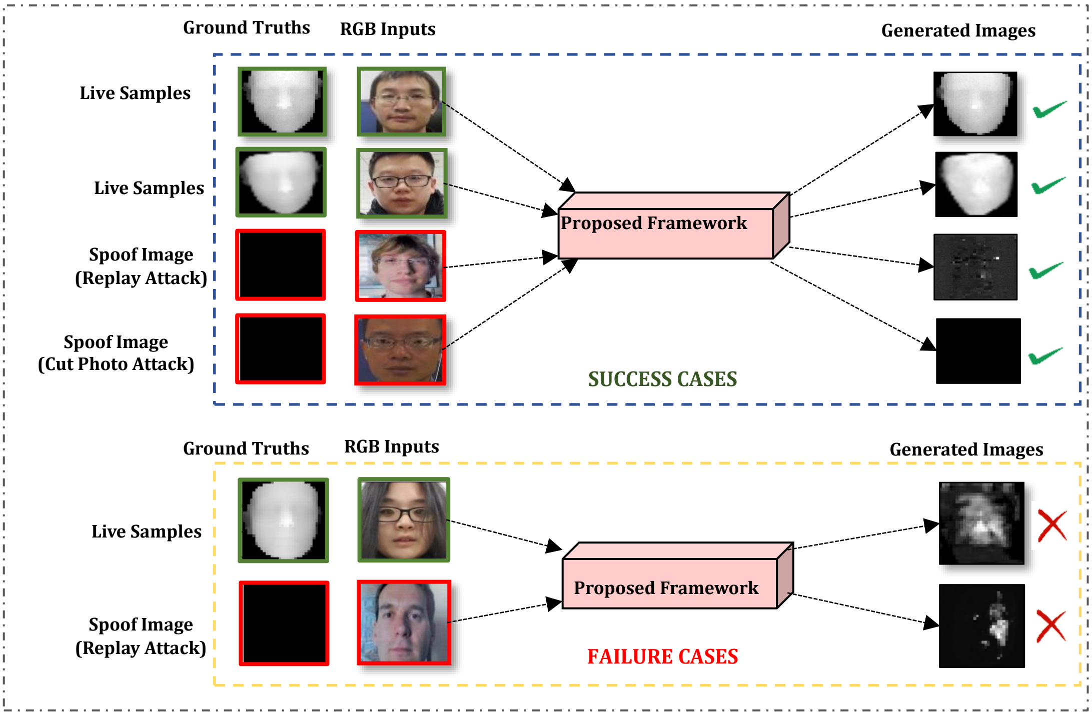


Figure 6.4. Comparison between Generated depth maps with Ground Truths. Row1: Input RGB images; Row2: Ground-Truth depth maps; and Row 3: Depth maps generated via proposed framework. The blue box highlights successful cases, while the red box displays failure cases.

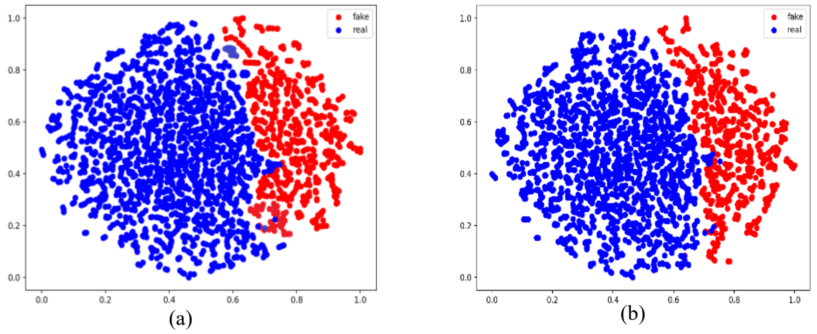


Figure 6.5 t-SNE feature distributions for (a) CASIA-FASD (C), (b) REPLAY-ATTACK (RA).

generator architectures to achieve robust and precise depth map generation.

6.3 Significant Outcomes

The outcomes of this chapter are summarized as follows:

- A novel GAN-based framework is proposed for FAS, integrating MSRCR with RGB inputs to improve input quality and provide discriminative visual cues crucial for spoof detection.

- The CBAM module is incorporated within the generative network to emphasize salient regions, thereby facilitating accurate and detailed depth map generation.
- Comprehensive intra- and cross-dataset evaluations are performed on four benchmark datasets— Replay-Attack, MSU-MFSD, CASIA-FASD, and OULU-NPU—demonstrating that the proposed work achieves performance comparable to other methods across diverse scenarios.

Although MSRCR enhances high-frequency information and CBAM improves saliency detection, both components predominantly focus on global enhancement. As a result, subtle spoof-specific artifacts may be underrepresented, leaving room for improvement in depth map precision. Depth maps generated in complex spoof scenarios occasionally lack fine-grained details, indicating the need for more advanced integration mechanisms to better exploit complementary features and further improve classification performance.

CASE 2

In Case 1, a GAN-based architecture was presented for face anti-spoofing (FAS), where RGB and MSRCR features were integrated alongside the CBAM to enhance saliency detection and depth-guided classification. While this framework demonstrated competitive performance and served as a strong foundation, certain limitations were observed, particularly in accurately capturing fine-grained spoof cues and generating high-quality depth maps under complex conditions.

To address these challenges, Case 2 introduces an improved framework, PolarSentinelGAN, which builds upon the strengths of Case 1. In this work, MSRCR is replaced with Multi-Scale Retinex with Color Preservation (MSRCP) to better retain high-frequency information critical for spoof detection. Furthermore, a Dual Polarized Self-Attention guided module (DPAttn) is proposed to intelligently evaluate and integrate RGB and MSRCP features. This attention mechanism enables the generative network to produce more accurate and detailed depth maps by prioritizing the most informative features. The refined features are then forwarded to an auxiliary classifier

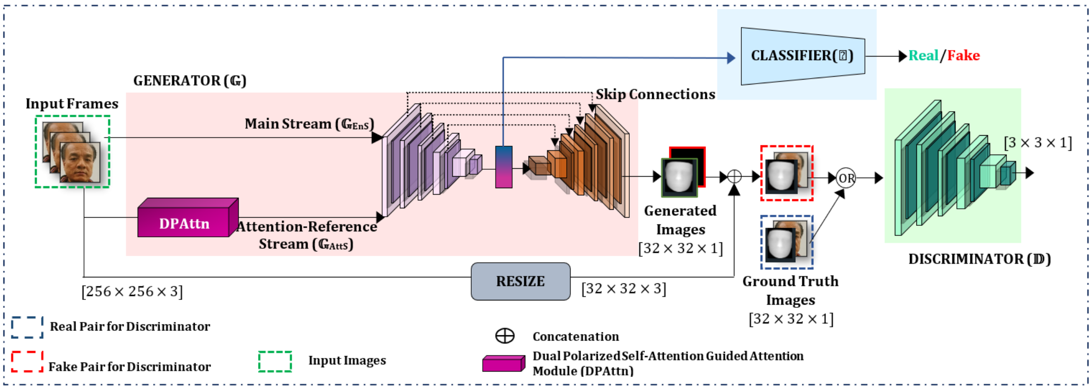


Figure 6.6 An outline of the proposed PolarSentinelGAN Framework. It consists of three key components: a generator (𝔾) responsible for generating depth maps from RGB images, a discriminator (𝔻) tasked with appraising the quality of the generator's outputs, and a classifier(ℂ) designed to distinguish genuine/spoofed face.

for robust live/spoof binary classification. Through this enhanced design, the limitations identified in Case 1 are effectively addressed, and a significant improvement in both depth estimation quality and anti-spoofing performance is demonstrated.

6.4 Methodology

In this section, the PolarSentinelGAN framework is introduced, as depicted in Figure 6.6. It is composed of three primary components: a generator (𝔾) through which depth maps are generated from RGB images, a discriminator (𝔻) by which the realism of generated outputs is evaluated, and a classifier (ℂ) by which live and spoof faces are distinguished. A dual-stream structure is employed in 𝔾, where RGB features are processed in the main stream, and complementary MSRCP features are incorporated through the Dual Polarized Self-Attention module (DPAttn) in the attention-reference stream. These modalities are selected due to their complementary properties—detailed textures are captured by RGB, while illumination-invariant, high-frequency features are retained by MSRCP. The encoded features are fused and passed to a decoder, where final depth maps are produced.

For 𝔻, a PatchGAN-based architecture [221] is adopted and implemented as a Fully Convolutional Network with a sequence of convolutional layers, each followed by batch normalization and a LeakyReLU activation. This design is favored over conventional binary classifiers to facilitate patch-level analysis for finer-grained authenticity evaluation. During training, 𝔻 receives input in the form of two [32 × 32] pairs: one real pair consisting of a ground truth depth map (D) and its corresponding RGB image (I_{RGB}), and one fake pair comprising a generated depth map $\{\text{Đ} = G(I)\}$ paired with RGB input. All images are resized from $[256 \times 256]$ to $[32 \times 32]$ dimensions prior to being processed. The primary objective of $\mathbb{G}$ is to generate depth maps from the input images (I) that are sufficiently realistic to deceive $\mathbb{D}$ into misclassifying them as genuine. Throughout this adversarial process, gradients from $\mathbb{D}$ are propagated to $\mathbb{G}$, enabling it to autonomously refine its generation capabilities. This iterative feedback mechanism allows the generator to progressively improve the realism and fidelity of the depth maps produced from RGB facial inputs.

A detailed discussion of each component of the proposed framework is presented in the following subsections.

6.4.1 Multi-Scale Retinex with Color Preservation (MSRCP)

Based on Retinex theory [217], three properties are fundamental for achieving perceptual image quality: dynamic range compression, color constancy under varying illumination, and accurate color interpretation. Dynamic range compression is accomplished through logarithmic image transformation, while color constancy is maintained by separating the illumination component. To ensure accurate color perception, the Multi-Scale Retinex (MSR) algorithm [217] is employed. MSR integrates outputs from the Single Scale Retinex [218] at small, medium, and large scales to balance dynamic range and color fidelity.

For a given input image $I(x, y)$, SSR is formulated as:

$$\mathbb{SSR} = log\ I(x,y) - log[I(x,y) * \mathcal{F}(x,y)] \tag{6.10}$$

where $*$ denotes the convolution operator, and $\mathcal{F}(x, y)$ is the center-surround function used to estimate the local average $I(x, y) * \mathcal{F}(x, y)$. A Gaussian filter ₲_σ, as suggested in [218], serves effectively as the surround function for illumination normalization:

$$\text{₲}_\sigma = Ke^{\frac{-(x^2+y^2)}{\sigma^2}} \tag{6.11}$$

here σ regulates spatial detail, color balance, and dynamic range, while the normalization constant K is chosen such that:

$$\iint \mathcal{F}(x,y)\, dxdy = 1 \tag{6.12}$$

Accordingly, the SSR for the i$^{\text{th}}$ color channel is expressed as:

$$\mathbb{SSR}_i(x,y) = log I_i(x,y) - log[(\mathbb{G}_\sigma * I_i)(x,y)] \quad (6.13)$$

where, $(\mathbb{G}_\sigma * I_i)(x,y)$ signifies the Gaussian blurring with standard deviations σ [σ_1=15, $\sigma_{2=}$ 80, $\sigma_{3=}$ 250 for small, medium, and large scale, respectively]. To jointly satisfy dynamic range compression and color consistency, MSR [217] is computed as a weighted summation of SSRs across scales:

$$\mathbb{MSR}_i(x,y) = \sum_{n=1}^{N} W_n\{log I_i(x,y) - \log{(I_i * \mathbb{G}_\sigma)(x,y)}\} \quad (6.14)$$

here, N represents the number of scales, and W_n denotes the corresponding weights. The result is then normalized to match the image's bit depth.

To mitigate spectral distortion and restore natural colors, Color Restoration (CR) is incorporated into the MSR output, forming MSRCR:

$$\mathbb{MSR}\not{\mathbb{C}}\mathbb{R}_i(x,y) = \mathbb{MSR}_i(x,y) \times \not{\mathbb{C}}\mathbb{R}_i(x,y) \quad (6.15)$$

The $\not{\mathbb{C}}\mathbb{R}_i(x,y)$ for ith channel at pixel position (x,y) is:

$$\not{\mathbb{C}}\mathbb{R}_\mathrm{i}(\mathrm{x},\mathrm{y}) = \beta[\log(\alpha \times \mathbb{I}(x,y))] \quad (6.16)$$

with $\mathbb{I}(x,y)$ represents chromaticity, computed as:

$$\mathbb{I}(x,y) = \frac{I_i(x,y)}{\sum_{j=1}^{ch} I_i(x,y)} \quad (6.17)$$

Substituting Eq. (8) into Eq. (7) yields:

$$\not{\mathbb{C}}\mathbb{R}_i(x,y) = \beta\{log(\alpha \times I_i(x,y)) - log(\sum_{j=1}^{ch} I_i(x,y))\} \quad (6.18)$$

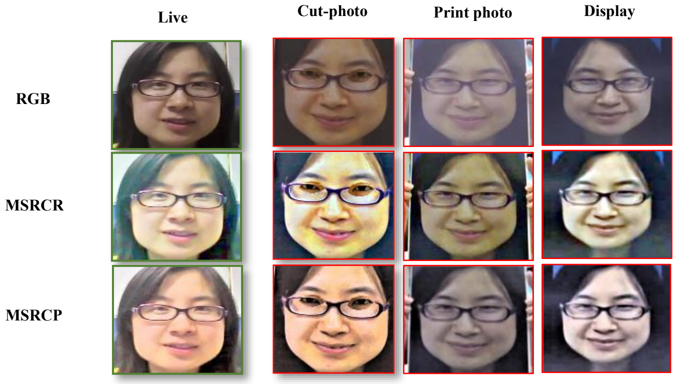
Here, α and β control non-linearity and total gain, while ch denotes the total number of channels. Incorporating gain (Ḡ) and offset (O) results in the complete MSRCR formulation:

$$\mathbb{MSR}\not{\mathbb{C}}\mathbb{R}_i = \bar{\mathrm{G}}\{\mathbb{MSR}_i(x,y) \times \not{\mathbb{C}}\mathbb{R}_i(x,y) - \mathrm{O}\} \quad (6.19)$$

With parameter values α=125, β=46, Ḡ=192 and O=30, MSRCR effectively enhances color. However, the independent application of CR to each channel can distort chromaticity, potentially introducing color shifts. To mitigate this, Multi-Scale Retinex (MSR) is implemented on the intensity channel [217], guarding against color inversion risks in the source image. This method preserves the original chromaticity and globally enhances color balance, upholding the image's authentic colors. As a result, the intensity image (Î), at pixel position (x,y) is calculated as:

$$\hat{\mathrm{I}}(x,y) = \frac{\sum_{j=1}^{ch} I_j(x,y)}{ch} \quad (6.20)$$

Figure 6.7 Pre-processed CASIA-FASD Images –green outlines (Live samples) and



red outlines (different types of attacks), arranged from Top to Down: RGB, MSRCR, and MSRCP.

MSR is then applied to Î, and a linear transformation normalizes the output to the [0-255] range:

$$\mathfrak{I}(x,y) = \sum_{n=1}^{N} W_n\{log\hat{\text{I}}(x,y) - log\big(\hat{\text{I}} * \text{₲}_{\sigma_n}\big)(x,y)\} \tag{6.21}$$

The enhancement is then proportionally propagated to each channel, maintaining both local and global color consistency. Thus, the MSRCP is defined for the ith channel at pixel position (x,y) as:

$$\mathbb{MSR}\text{₵}\mathbb{P}_i(x,y) = \mathcal{J}_i(x,y) \times \mathfrak{B}(x,y) \tag{6.22}$$

where the brightness scaling function $\mathfrak{B}(x,y)$ is given by:

$$\mathfrak{B}(x,y) = min(\frac{255}{max(I_j(x,y))}, \frac{\mathfrak{I}(x,y)}{\hat{\text{I}}(x,y)}) \quad j\epsilon(1..ch) \tag{6.23}$$

Unlike MSRCR, which operates independently on each color channel and can risk chromaticity distortion. MSRCP applies enhancement globally through the intensity channel, leading to more natural and visually coherent color representation. While the performance of both methods is influenced by lighting conditions, MSRCP has shown superior robustness under white or colored illumination, making it particularly suitable for complex real-world scenarios. Figure 6.7 displays pre-processed samples from the CASIA-FASD dataset, illustrating genuine and spoof samples with RGB, MSRCR, and MSRCP representations. In the generator, MSRCP is employed as detailed in Section IV. The ablation study presented therein validates its superior performance under colored lighting conditions, demonstrating a clear advantage over traditional MSRCR.

6.4.2 Dual Polarized Self-Attention Guided Module (DPAttn)

The attention-reference stream integrates the Dual Polarized Self-Attention Guided Module (DPAttn), which exploits the complementary information present in both RGB (I_{RGB}) and MSRCP (I_{MSRCP}) representations. By aligning and reinforcing the correspondence between salient features and regions of interest, this module enables more effective feature discrimination. Incorporating MSRCP alongside RGB within the generative network proves especially beneficial for GAN-based feature mining, as it enhances the architecture's ability to capture subtle spoofing cues and extract illumination-invariant live/spoof features. It fully exploits the synergy between I_{RGB} and I_{MSRCP}, the proposed approach first applies a Polarized Self-Attention (PSA) mechanism, which selectively emphasizes informative regions in both modalities. The refined feature maps produced by PSA are subsequently passed through a Weighted Fusion Block (WFB), which further enhances the representational quality through adaptive fusion, resulting in a more robust and semantically enriched feature representation.

Polarized Self-Attention (PSA). The PSA [228] serves as a specialized self-attention module designed to address a common challenge faced by traditional deep CNNs– the loss of high-resolution information during pooling or downsampling. Unlike other attention mechanisms, PSA effectively preserves distinct information in the orthogonal direction. This preservation is achieved through a polarized filtering mechanism that segregates channel-specific and spatial specific information, allowing for more precise attention weighting. Furthermore, PSA's dual-block architecture, with one block dedicated to channel-only attention (PS-CA) and the other to spatial-only attention (PS-SA) provides fine-grained control over the attention process. This structure leads to improved performance in tasks that demand precise localization and feature extraction. The unique nonlinear composition strategy employed by PSA seamlessly merging softmax-sigmoid compositions within the PS-CA and PS-SA branches, contributes to harmonizing the output distribution characteristics with the nuanced requirements of fine-grained regression tasks. This combination of features distinguishes PSA from other attention mechanisms, making it a powerful tool applicable across a wide range of computer vision tasks.

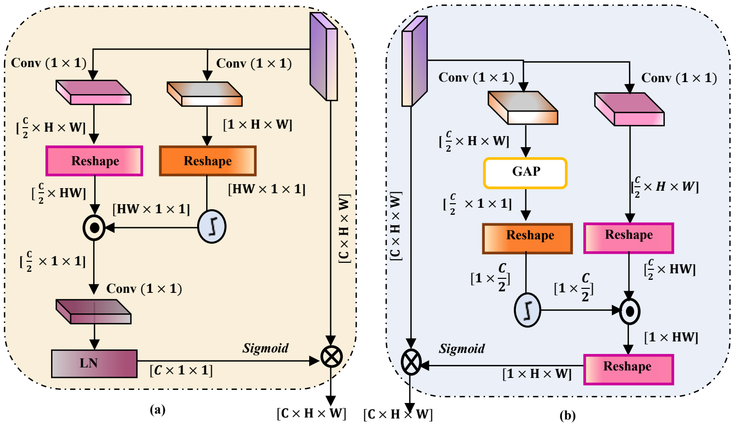


Figure 6.8 (a) Polarised Self-Channel Attention (PS-CA), (b) Polarised Self-Spatial Attention (PS-SA).

In our framework, PS-SA is applied to I_{MSRCP}, while PS-CA is employed for the I_{RGB}. This approach allows us to capture long-range contextual information in both spatial and channel dimensions, respectively. As illustrated in Figure 6.8, given a feature representation $F\epsilon\ \mathbb{R}^{C\times H\times W}$, we instantiate the PSA mechanism as follows:

- ***Channel-only branch (PS-CA):*** This process generates the channel attention map $C_{maps}\epsilon\mathbb{R}^{C\times 1\times 1}$, represented by:

  $$C_{maps} = F_{SG}[\,conv_{1x1|\Theta_1}((\sigma_1(conv_{1x1}(F))\odot F_{SM}(\sigma_2(conv_{1x1}(F))))] \quad (6.24)$$

  where Θ_1 is an intermediate parameter for these channel convolutions. σ1 and σ2 are two tensors reshape operators, and ⊙ is the matrix dot-product operation. F_{SG} and F_{SM} are the sigmoid and softmax operators respectively.

- ***Spatial-only branch (PS-SA):*** The generated spatial attention map $S_{maps}\epsilon\mathbb{R}^{C\times 1\times 1}$, represented by:

  $$S_{maps} = F_{SG}[\,\sigma_3(F_{SM}(\sigma_1(GAP(conv_{1x1}(F))))\odot\sigma_2(conv_{1x1}(F)))] \quad (6.25)$$

  where σ_1, σ_2 and σ_3 are three tensor reshape operators, and GAP is a global average pooling operator.

Weighted Fusion Block (WFB). Applying PS-CA to I_{RGB} and PS-SA to I_{MSRCP} generates feature maps C_{maps} and S_{maps}, respectively. To thoroughly explore their interaction, we devised a Weighted Fusion Block (WFB) within the DPAttn module. It fine-tunes the impact of C_{maps} and S_{maps}, refining the output through an attention

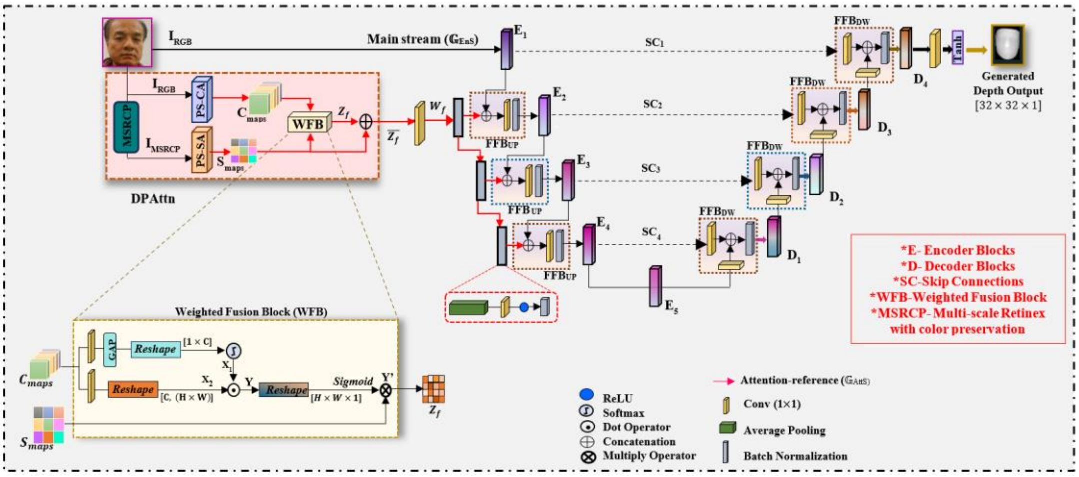


Figure 6.9 The Generator Architecture of our proposed framework PolarSentinelGAN. The generator in PolarSentinelGAN features a dual-stream design to enhance depth map reconstructions. The first stream, called Main Stream, 𝔾EnS utilizes encoder blocks to process RGB images as input. Simultaneously, the Attention- Reference Stream, 𝔾AttS operates as a second stream, employing a dual polarized self-attention guided attention module (DPAttn) to effectively process both RGB and MSRCP images.

mechanism. It optimizes effectiveness by selectively enhancing informative features. As shown in Figure 6.9, WFB calibrates the output of each subnetwork, aligning the overall output with the significance of the feature maps. Specifically, for C_{maps}, it employs an attentional mechanism, as expressed by following equations:

$$X_1 = F_{SM}\left(\sigma_4\left(GAP\left(conv_{1x1}\left(C_{maps}\right)\right)\right)\right) \tag{6.26}$$

$$X_2 = \sigma_5\left(conv_{1\times1}\left(C_{maps}\right)\right) \tag{6.27}$$

$$Y = X_1 \odot X_2 \tag{6.28}$$

$$Y' = F_{SG}\left(\sigma_6(Y)\right) \tag{6.29}$$

here σ_4, σ_5 and σ_6 are three tensor reshape operators, and “$\odot$” signifies the dot product. To harness the aggregated information Y, we introduce gating mechanism using a sigmoid function, F_{SG}. This produces a soft attention map, Y', indicating the relative weights of different feature regions. Y’ is then multiplied with the S_{maps} emphasizing important features while suppressing less important ones. This approach encourages the network to learn intricate relationships within encoded representations, ultimately resulting in the output feature Z_f, expressed as follows:

$$Z_f = Y' \otimes S_{maps} \tag{6.30}$$

The output of the WFB is then concatenated with S_{maps} to generate the final refined output $\overline{Z_f}$ expressed as:

$$\overline{Z_f} = Z_f \oplus S_{maps} \tag{6.31}$$

This enhanced feature map guides the generative network through each encoder stage to produce depth maps.

6.4.3 Generative Network

The generator $\mathbb{G}$ is designed with an encoder-decoder structure as shown in Figure 6.9. Within this architecture, the input images are transformed into compact, lower-dimensional representations by the encoder, allowing the decoder to reconstruct depth maps using only these encoded features. The success of this process is determined by the encoder's capacity to extract essential features necessary for depth reconstruction. To achieve a balance between model capacity and efficiency, EfficientNetB4 [185] has been selected as the backbone of the generator. To promote the extraction of multi-scale features, skip connections are integrated to form a U-Net network. These connections allow encoder representations to be seamlessly merged with their corresponding up sampling counterparts in the decoder. As a result, enhanced gradient flow is enabled and the vanishing gradient problem—especially in earlier layers—is mitigated.

A dual-stream strategy is employed by $\mathbb{G}$ to generate meaningful depth maps, comprising the main stream and the attention-reference stream. The main stream, $\mathbb{G}_{\text{EnS}}$ processes the I_{RGB} input through encoder blocks. In parallel, the attention-reference stream, $\mathbb{G}_{\text{AttS}}$ utilizes the Dual Polarized Self-Attention guided module (DPAttn), which operates on both I_{RGB} and I_{MSRCP}. As shown in Figure 6.9, DPAttn evaluates the saliency of features across channels and spatial locations from the two modalities. Specifically, the PS-SA module is applied to I_{MSRCP} generating $\text{S}_{\text{maps by}}$ aggregating global contextual information across pixels. Concurrently, PS-CA is applied to I_{RGB}, yielding C_{maps} by capturing inter-channel dependencies. These two attention maps are fused through the Weighted Fusion Block (WFB), resulting in enhanced face features

$\overline{Z_f}$. A convolutional block is subsequently applied to $\overline{Z_f}$ to produce W_f which undergoes progressive downsampling through dedicated layers.

For a given feature input $\text{Ƒ}\epsilon\mathbb{R}^{C_1\times H_1\times W_1}$, the nth down sampled output ρ ∈ $\mathbb{R}^{(C_2\times H_2\times W_2)}$ is computed as:

$$\rho_n = BN\left(relu\left(conv_{1\times 1}\left(avgpool(\text{Ƒ}_{n-1})\right)\right)\right) \tag{6.32}$$

where BN denotes Batch Normalization. This downsampling ensures dimensional alignment of W_f with the encoder block E_i to which it is passed as a reference input within the $\mathbb{G}$EnS stream. Within the $\mathbb{G}$EnS stream, the output from the previous encoder block, E_{i-1}^{out} is fused with the corresponding attention stream input ρ_i using Feed Forward Blocks (FFBs). The output of each encoder block E_i is given by:

$$E_i^{out} = E_i[FFB(E_{i-1}^{out}, \rho_i)] \quad i\epsilon[1,5] \tag{6.33}$$

Through this consistent application of FFB, seamless feature convergence is facilitated, allowing the attention-reference stream to guide the main stream and ensuring the discriminative features extraction passed to the decoder.

In the decoder, each block receives inputs from both the preceding decoder stage and its corresponding encoder block via skip connections. These inputs are aligned through convolution, concatenation, and batch normalization. As illustrated in Figure 6.9, the FFB structure plays a crucial role in enabling information fusion in both encoder and decoder components.

Forward Feeded Block (FFB). As discussed, the FFB is designed to integrate complementary feature representations from $\mathbb{G}$EnS and $\mathbb{G}$AttS. It facilitates the smooth transfer of semantic, spatial, and contextual details, thus improving the framework's overall comprehension capabilities.

Loss Function. For a given input I, the GAN objective function is defined as:

$$\mathbb{L}_{GAN}(\mathbb{G},\mathbb{D}) = \mathbb{E}_{I,\mathrm{D}}[log\mathbb{D}(I,D)] + \mathbb{E}_I[log(1-\mathbb{D}(I,\text{Đ})] \tag{6.34}$$

here, D represents ground truth depth maps, and $\text{Đ} = \mathbb{G}(I)$. s part of a zero-sum game, the optimization of $\mathbb{G}$ and $\mathbb{D}$ is performed alternately. During the discriminator update, only $\mathbb{D}$ is updated by minimizing:

$${}^{min}_{\mathbb{D}}\mathbb{L}_{GAN}^{\mathbb{D}} = -\mathbb{E}_{I,D}[log\mathbb{D}(I,D)] - \mathbb{E}_I[log(1-\mathbb{D}(I,\text{Đ})] \tag{6.35}$$

In this step, the discriminator learns to distinguish real from generated depth maps. For the generator, the first term of Eq. (25) is excluded as it is independent of $\mathbb{G}$. Thus, the objective for $\mathbb{G}$ becomes:

$$\mathbb{L}_{GAN}^{\mathbb{G}} = -\mathbb{E}_I[log(1 - \mathbb{D}(I, \text{Đ})] \tag{6.36}$$

To further enhance the realism of generated depth images, an L1 reconstruction loss is introduced:

$$\mathbb{L}_{L1}(\mathbb{G}) = \mathbb{E}_{I,D}[||D - \text{Đ}||_1] \tag{6.37}$$

This loss does not influence $\mathbb{D}$, but guides $\mathbb{G}$ to produce outputs closer to ground truth. The final training objective of the PolarSentinelGAN framework is formulated as:

$$\mathbb{L}_{DGA-GAN}(\mathbb{G}, \mathbb{D}) = arg_{\mathbb{G}}^{min\,max}{}_{\mathbb{D}}\mathbb{L}_{GAN}(\mathbb{G}, \mathbb{D}) + \lambda\mathbb{L}_{L1}(\mathbb{G}) \tag{6.38}$$

here λ is a balancing parameter. This optimization ensures effective training and enables the generation of high-quality depth maps.

6.4.4 Classification

The classification task is accomplished using the latent variables extracted from the encoder of $\mathbb{G}$, which are utilized as the final output representation for FAS. The distinction between live and spoofed faces is performed based on features that encapsulate facial depth information. These features are downsampled and fused using (3×3) convolutional modules. Subsequently, a fully connected layer is applied to derive the classification label.

During GAN training, the encoder is fine-tuned to enable the effective fusion of RGB and depth-based features, thereby enhancing the architecture's ability to generalize across varying spoofing attacks. The classifier is optimized using the binary cross-entropy loss, defined as:

$$\mathbb{L}_C = -(y\, log(p) + (1 - y)\, log(1 - p)) \tag{6.39}$$

where y represents the ground truth label and p depicts the predicted probability.

6.5 Experimental Analysis

In this section, we detail experimental setup, results, benchmark comparisons for each dataset, and an ablation study to analyze performance variations.

6.5.1 Implementation Details

Data Preprocessing. In line with prior research practices, we initiated pre-processing with frame sampling and face alignment. For video-sourced evaluation benchmarks, we systematically extracted frames at every 10th frame interval, adhering to the respective dataset guidelines. Viola-Jones algorithm [205] initially facilitated face detection and Region of Interest (ROI) extraction, standardizing detected faces to$[256 \times 256 \times 3]$ as RGB inputs. However, Viola-Jones algorithm limitations affected 4.3% of frames from the RA dataset, 3.7% from C, and 7% from M dataset. In response, we adopted the MTCNN [170] as an alternative for face extraction. To generate ground-truth depth maps, we embraced a dense face alignment strategy (i.e., PRNet [229]), yielding $[32 \times 32]$ dimensions for actual faces and spoof depth maps are set to zeros. Random horizontal flips and data augmentation normalization techniques were applied to enhance dataset diversity and model robustness.

Training Setup. Implemented in Keras, the proposed framework runs in the Google Colab Pro environment, utilizing an Nvidia T4 GPU equipped with RAM of 16 GB. EfficientNetB4 [185] serving as the backbone of generator, is initialized with a pre-trained ImageNet model. The newly incorporated layers adopt the “HeUniform” initialization approach. We optimize the architecture using the Adam optimizer, initializing $1e-3$ as the learning rate, and setting 16 as batch size. Loss function configuration sets λ_{GAN} and λ_{L1} to 1 and 100, respectively, following [104]. Each training epoch involves random shuffling and flipping of images to augment dataset diversity. During inference, the framework achieves a processing time of approximately 0.773 seconds per batch of 16 images, demonstrating its efficiency in handling real-time applications.

6.5.2 Comparative Analysis with other-state-of-the-arts

To evaluate the effectiveness of our approach, a comprehensive comparison with other methods was conducted, as depicted in Tables 6.7- 6.12. This included both intra- and cross-dataset testing across various datasets.

Intra-testing Testing. We rigorously conducted intra-dataset testing on CASIA-FASD(C), ROSE-Youtu (RY), Replay-Attack (RA) and OULU-NPU (O) following prescribed evaluation protocols. In these experiments, our approach was benchmarked against several state-of-the-art methods, including GAN-based techniques such as

Table 6.7 Intra-dataset testing comparison within CASIA-FASD(C) (%)

Methods	EER
Color Texture [24]	6.20
CNN [46]	4.64
Patch and Depth [73]	2.67
Attention [53]	3.14
FARCNN [151]	2.35
DTN [104]	1.34
Zhang et al. [110]	1.17
Ours	**1.15**

Table 6.8 Intra-dataset testing comparison Replay-Attack (RA) (%)

Methods	EER	HTER
Color Texture [24]	0.4	2.9
Patch and Depth [73]	0.79	0.72
FARCNN [151]	0.06	0.18
Attention [53]	0.13	0.25
DTN [104]	0.06	0.02
Zhang et al. [110]	0.09	0.22
Ours	**0.04**	**0.02**

Table 6.9 Intra-dataset testing comparison ROSE-Youtu (%)

Methods	HTER
Ensemble of classifiers [230]	9.3
FASNeT [231]	8.57
Fatemifar et al. [232]	6.34
Alassafi et al. [233]	4.92
Our	**1.25**

DTN [104], Zhang et al. [110], CSM-GAN [234], and Liu et. al [208]. Table 6.7 presents the EER values for the C dataset, while Table 6.8 highlights the EER and HTER metrics for the RA dataset. As shown in both Tables, our approach consistently outperforms other methods, including GAN-based ones [110] [104]. Similarly, Table 6.9 provides HTER-based evaluation results for the ROSE-Youtu dataset, where our method also demonstrates superior capability compared to other techniques. Table 6.10 showcases extensive experiments on the O dataset across all four prescribed protocols. In protocols 1 and 2, our method slightly underperforms in terms of APCER and BPCER compared to CSM-GAN [234] and Liu et. al [208]. However, on average, our work surpasses other works, resulting in impressively low ACER values of 0.27% and 1.05%, respectively. Under protocol 3, our approach performs below CSM-GAN [234], and Liu et. al [208], with a minimal difference. Notably, under protocol 4, our approach achieves the best performance among all methods, yielding 2.1±3.3% of

Table 6.10 Intra-dataset testing comparison OULU-NPU (O) (%).

Protocols	Methods	APCER	BPCER	ACER
1	Auxiliary [93]	1.6	1.6	1.6
	SGTD [81]	2	0	1
	STDN [106]	0.8	1.3	1.1
	CDCN [70]	0.4	1.7	1.1
	CSM-GAN [234]	0.14	0.56	0.35
	Liu et. al [208]	0.6	0.0	0.3
	DTN [104]	0.78	1.06	0.92
	Zhang et al. [110]	0.63	0.80	0.72
	Ours	**0.25**	**0.3**	**0.27**
2	Auxiliary [93]	2.7	2.7	2.7
	SGTD [81]	2.5	1.3	1.9
	STDN [106]	2.3	1.6	1.9
	CDCN [70]	1.5	1.4	1.5
	CSM-GAN [234]	0.69	1.67	1.18
	Liu et. al [208]	0.7	1.4	1.1
	DTN [104]	3.84	2.11	2.88
	Zhang et al. [110]	2.53	1.36	1.95
	Ours	**1.3**	**0.8**	**1.05**
3	Auxiliary [93]	2.7±1.3	3.1±1.7	2.9±1.5
	SGTD [81]	3.2±2.0	2.2±1.4	2.7±0.6
	STDN [106]	1.6±1.6	4.0±5.4	2.8±3.3
	CDCN [70]	2.4±1.3	2.2±2.0	2.3±1.4
	CSM-GAN [234]	0.50±0.97	2.83±1.38	1.67±1.05
	Liu et. al [208]	1.5±1.3	1.4±1.3	1.5±1.1
	DTN [104]	1.9±1.6	3.8±6.4	2.8±2.7
	Zhang et al. [110]	1.7±1.4	2.7±4.3	2.2±3.0
	Ours	**1.3±1.7**	**2.1±1.0**	**1.7±1.1**
4	Auxiliary [93]	9.3±5.6	10.4±6.0	9.5±6.0
	SGTD [81]	6.7±7.5	3.3±4.1	5.0±2.2
	STDN [106]	2.3±3.6	5.2±5.4	3.8±4.2
	CDCN [70]	4.6±4.6	9.2±8.0	6.9±2.9
	CSM-GAN [234]	2.22±1.77	8.29±4.18	5.26±2.88
	Liu et. al [208]	4.2±3.0	1.7±2.6	3.0±1.9
	DTN [104]	4.0±4.1	3.0±4.9	3.5±2.4
	Zhang et al. [110]	2.1±4.5	5.7±4.9	3.9±3.2
	Ours	**1.8±2.2**	**2.5±4.3**	**2.1±3.3**

ACER. These results affirm that our work meets state-of-the-art criteria, especially under complex scenarios involving unknown attacks, device disparities, and varying lighting conditions.

Cross-Dataset Testing. Evaluating the cross-dataset adaptability of FAS frameworks is vital for ensuring their effectiveness in practical applications. Therefore, we conducted cross-dataset evaluation to assess the generalization ability of our framework, utilizing datasets C, RA, and RY. In these experiments, the proposed face

Table 6.11 Comparative Cross-dataset analysis of CASIA-FASD, Replay-Attack and ROSE-Youtu (HTER %).

Methods	Train- RA &Test-C	Train- RA &Test- RY	Train- C &Test- RA	Train- C &Test- RY	Train- RY &Test-C	Train- RY &Test-RA
Auxiliary [93]	28.4	-	27.9	-	-	-
FARCNN [151]	29.4	-	26.0	-	-	-
Tzeng et al. [235]	49.8	50.0	41.8	31.4	28.7	34.6
Attention [53]	33.4	-	30.0	-	-	-
CSM-GAN [234]	23.4	-	26.8	-	-	-
Liu et. al [208]	26.7	-	22.0	-	-	-
Li et al. [111]	12.3	40.1	39.3	31.6	30.1	38.8
Wang et al. [112]	41.5	41.7	17.5	29.4	34.1	30.3
DTN [104]	22.98	-	16.64	-	-	-
Zhang et al. [110]	21.57	-	25.73	-	-	-
Alassafi et al. [233]	23.09	33.27	28.89	32.03	29.78	8.43
Ours	**20.62**	**19.25**	**15.56**	**23.96**	**21.16**	**8.05**

PAD framework was trained on the RA training dataset and evaluated on RY and C datasets. Similarly, the framework was trained on individual datasets and tested on others, with HTER (%) values summarized in Table 6.11. As observed from the results, our approach consistently demonstrates robustness and versatility in cross-database testing scenarios, achieving the best performance among the evaluated methods.

To further validate our approach in handling intricate scenarios, we adopted a protocol involving four datasets (CASIA-FASD(C), OULU-NPU(O), Replay-Attack (RA), and MSU-MFSD(M). Each served as the test set in one of four, with the remaining three as the training set, resulting in four distinct test cases: Test Case 1- O&C&RA→ M, Test Case 2- O&M&RA → C, Test Case 3-O&C&M→ RA, and Test Case 4- RA&C&M→ O. Table 6.12 shows that for former two test cases (1 and 2), our approach outperformed other methods. However, in the case of Test Case 3, our proposed approach exhibited a slight lag in performance compared to MADGG [210], DR-UDA [113],and DTN [104], while it performed well in Test Case 4. Despite a minor lag in Test Case 3, our approach demonstrated the ability to generate descriptive depth maps and features, enabling the classifier to distinguish between live and spoof samples. While methods like GDA [236], EPCR [237], and DTDA [238] achieve better results overall, their superior performance is driven by fundamentally different and more complex multi-stage strategies, which inherently increase inference time and

Table 6.12 Comparative Cross-Dataset analysis on four datasets (%).
Test Case 1: C&RA&O→ M; Test Case 2: M&RA&O→C; Test Case 3: C&M&O→RA; Test Case 4: C&RA&M→O

Methods	Test Case 1		Test Case 2		Test Case 3		Test Case 4	
	HTER	AUC	HTER	AUC	HTER	AUC	HTER	AUC
Auxiliary (Depth Only) [93]	22.72	85.88	33.52	73.15	29.14	71.69	30.17	77.61
Binary CNN [46]	29.25	82.87	34.88	71.94	34.47	65.88	29.61	77.54
DR-UDA [113]	16.1	-	22.2	-	22.7	-	24.7	-
MADDG [210]	17.69	88.06	24.5	84.51	22.19	84.99	27.98	80.02
DTN [104]	19.40	86.87	22.03	87.71	21.43	88.81	18.26	89.40
GDA [236]	9.2	98.0	12.2	93.0	10.0	96.0	14.4	92.6
EPCR [237]	12.5	95.3	18.9	89.7	14.0	92.4	17.9	90.9
DTDA [238]	5.71	98.03	6.67	97.27	13.12	92.24	13.13	94.24
Ours	**15.8**	**91.5**	**20.15**	**85.4**	**23.11**	**84.1**	**16.53**	**90.43**

computational complexity. In contrast, PolarSentinelGAN adopts a streamlined and efficient design, leveraging GAN-based depth map generation and a polarized attention module to deliver competitive results without compromising simplicity or speed. By prioritizing real-time applicability and resource efficiency, PolarSentinelGAN achieves a robust balance between accuracy and operational efficiency, which makes it highly advantageous for practical deployment. Furthermore, cross-dataset experiments remain a valuable avenue for future exploration to further highlight the model's strong generalization capabilities.

6.5.3 Ablation Study

In the following section, we perform ablation studies to highlight the significance of selecting individual components in our framework. These ablation studies are exclusively performed using the OULU-NPU protocol 2 as our testing benchmark.

Effect of different Generator architecture. We explore the impact of various base networks within our proposed framework (see Table 6.13). Our initial experiments involved a basic encoder-decoder architecture, which gives a relatively high ACER i.e., 10.2%. This model faced challenges in accurately generating depth maps for genuine faces due to its bottleneck design, necessitating information to traverse all layers. Subsequently, we adopted the U-Net network with skip connections, substantially improving performance and reducing to 8.65% ACER. This architectural.

Table 6.13 Ablation study on OULU-NPU protocol 2(%).

Backbone	**Input**	**APCER**	**BPCER**	**ACER**
Simple Encoder-Decoder (w/o skip connections)	RGB	11.5	8.9	10.2
U-Net	RGB	9.6	7.7	8.65
	RGB+MSRCR	5.4	4.8	5.1
	RGB+MSRCP	4.5	4.4	4.45
U-Net + CBAM	RGB+MSRCP	3.8	3.6	3.7
U-Net + PSA (with simple concatenation)	RGB+MSRCP	3.0	3.2	3.1
U-Net + DPAttn (PSA+WFB)	RGB+MSRCP	2.1	2.5	2.3
U-Net + DPAttn+ FFB$_{DW}$	RGB+MSRCP	1.5	0.9	1.2
U-Net + DPAttn+ FFB$_{DW}$ + FFB$_{UP}$	**RGB+MSRCP**	**1.3**	**0.8**	**1.05**

adjustment effectively retained features, resulting in an overall enhancement

Effect of Input variants. Table 6.13 summarizes the pivotal role of input types in our framework. Initially, employing standard RGB input resulted in 8.65% ACER on the U-Net architecture, revealing limitations in creating accurate depth maps using RGB alone. To enhance depth map generation, we introduced the Multi-Scale Retinex (MSR) technique, comprising both MSRCR and MSRCP. Combining MSRCR and MSRCP with RGB input enhanced the extraction of distinctive features, particularly in challenging lighting conditions. Our findings showed a notable reduction in ACER to 5.1 % with RGB and MSRCR, reaching the lowest ACER of 4.45% with RGB and MSRCP. This highlights the crucial role of input type selection in enhancing the overall effectiveness of the framework, especially in generating near-accurate depth maps for real faces.

Effect of DPAttn. Initially, the U-Net baseline network with led to confusion in distinguishing genuine and spoof faces, thereby negatively impacting depth map generation quality. To overcome this, we introduced the 𝔾AttS stream alongside the 𝔾EnS stream, initially employing the CBAM attention mechanism, provided limited improvement. To refine facial features and enhance the generator's output, we then adopted DPAttn. As shown in Table 6.13, the adoption of DPAttn significantly lowered ACER to 1.3%, highlighting substantial depth generation quality enhancement and reinforcing the effectiveness of employing a U-Net with an attention block.

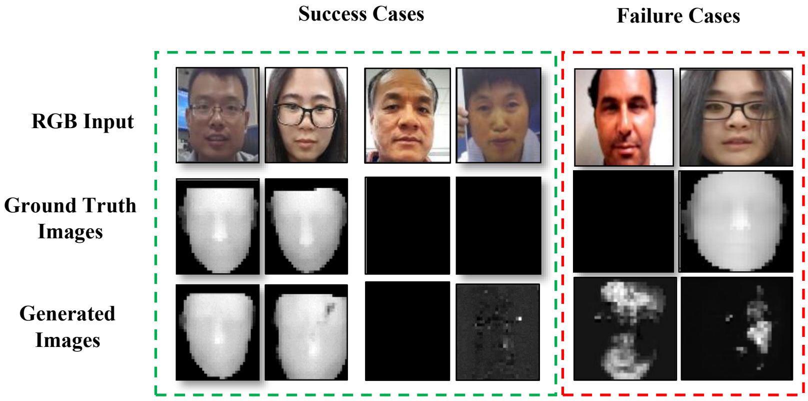


Figure 6.10 Comparison between generated depth images and Ground Truths for given RGB inputs. First Row: RGB domain input images; Second Row: Ground-Truth depth maps; and Third Row: Depth maps generated. Green box-Columns 1 to 6 display success cases, while red box-columns 7 to 9 depict failure cases.

Effect of Forward Feeded block. To further boost the framework's performance, we integrated FFBs at each encoder stage, effectively combining relevant features from previous encoder blocks with the 𝔾AttS stream. The use of FFBDW significantly reduced ACER, and extending FFBUP at every decoder stage further improved feature fusion, ultimately achieving an impressive ACER of 1.05% as shown in Table 6.13. These results provide strong evidence of our proposed framework's effectiveness.

6.5.4 Visualization and Analysis

In Figure 6.10, we illustrate the performance of our framework in generating depth maps from input RGB images. The first row displays the original RGB images, the second row shows the ground-truth images, and the third row presents the depth maps produced by our framework. For live samples, our framework generates depth maps that closely resembles the ground-truth, with only minor deviations in fine details. In contrast, for spoofed samples, it primarily produces zero-depth maps, though occasional outputs exhibit noise-like patterns. Figure 6.10 also includes a subset of misclassified samples. Despite the image being spoofed face, our framework produce image that closely resemble facial depth map. Another misclassification example shows the framework's failure to accurately generate a depth map for a given live sample, revealing a fundamental issue causing classification errors.

For a more intuitive understanding of our framework decision-making process,

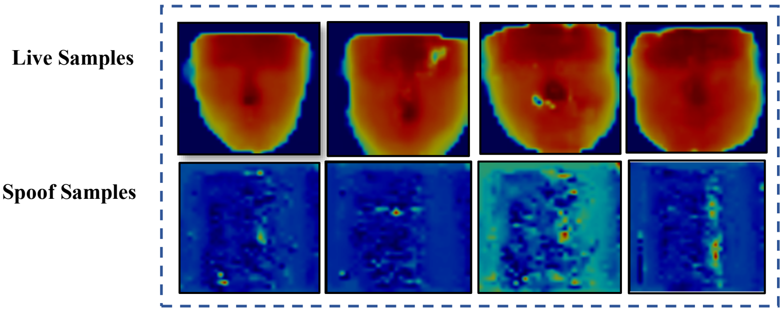


Figure 6.11 Visualizing Grad-CAM attention maps for various test samples, the first row demonstrates focused attention on live faces, while the second row showcases attention patterns learned for spoofed faces.

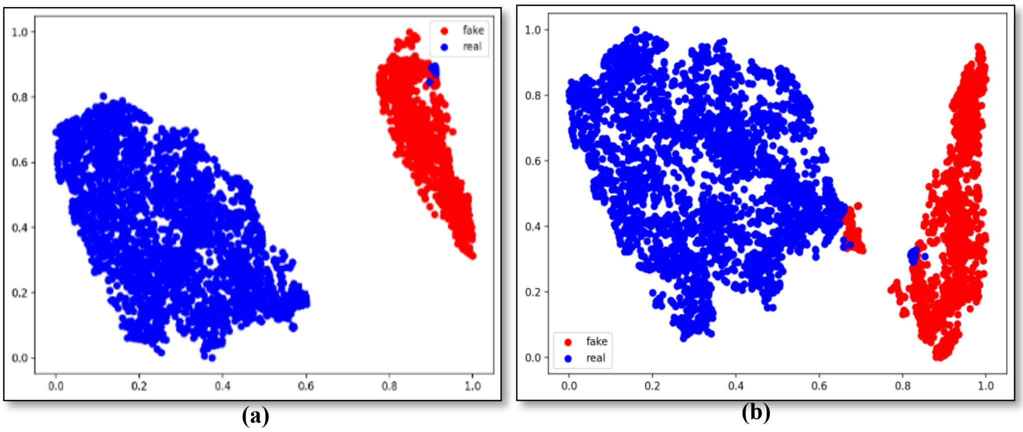


Figure 6.12 t-SNE feature distribution for (a) CASIA-FASD, (b) REPLAY-ATTACK.

Figure 6.11 provides visualizations using the Grad-CAM (Gradient- weighted Class Activation Mapping) [213] technique. The first and second rows of Grad-CAM maps generated from live and spoofing faces, respectively visualize class attention maps. Notably, attention for live faces is primarily concentrated on the facial region, reflecting the origin of most depth information. Conversely, attention distribution for spoofing faces appears random, as they lack meaningful depth information.

The t-SNE [214] visualization algorithm is employed to illustrate the distinctive strength of the CNN-derived features for FAS. Figure 6.12 (a) presents the feature representation of the CASIA-FASD dataset with a model trained on CASIA-FASD, while Figure 6.12 (b) illustrates the feature distributions of the Replay-Attack dataset with model trained on Replay-Attack. These visualizations demonstrate that the proposed framework effectively transforms RGB inputs into the depth domain, producing consistent representation for actual and spoofed data across databases. This transformation notably enhances the generalization capacity of the framework.

6.6 Significant Outcomes

The outcomes of this chapter are summarized as follows:

- PolarSentinelGAN extends a prior GAN-based FAS framework by replacing MSRCR with MSRCP, enabling better preservation of high-frequency spoof cues and improved depth map generation under complex conditions.
- A novel Dual Polarized Self-Attention Module (DPAttn) adaptively fuses RGB and MSRCP features, guiding the generator to focus on modality-specific discriminative information.
- Feed Forward Blocks (FFBs) are embedded within the encoder-decoder to facilitate seamless integration of attention-refined features and enhance depth map coherence.
- The framework demonstrates state-of-the-art performance across standard intra- and cross-dataset benchmarks, validating its strong generalization capabilities.
- Ablation studies confirm the effectiveness of each component—MSRCP enhancement, DPAttn, and FFBs—in boosting detection accuracy and robustness.

CHAPTER 7

CONCLUSION, FUTURE SCOPE, AND SOCIAL IMPACT

This thesis addresses robust and automatic face anti-spoofing using advanced deep learning frameworks, driven by the need to protect biometric systems from presentation attacks. Despite its significance, face anti-spoofing remains challenging due to variations in lighting, backgrounds, spoofing materials, and image quality, along with poor generalization to unseen attacks. Capturing both spatial and temporal cues adds further complexity. To tackle these challenges, we propose four novel approaches that aim to enhance feature representation, improve generalization capability, and ensure computational efficiency. The proposed methods are scalable, real-time, and explainable, offering strong potential for deployment in real-world biometric authentication systems. A comprehensive evaluation of each framework demonstrates its effectiveness, as summarized in Section 7.1, followed by the discussion on future directions and societal impact in Section 7.2.

7.1 Summary of the Work Done in the Thesis

Four major approaches of the face anti-spoofing based on traditional handcrafted features and deep features are presented and these approaches are as follows:

1. The first proposed solution presents an efficient RGB based FAS framework that integrates multi-level ELBP with a modified Xception network enhanced by squeeze-and-excitation mechanisms. This approach improves feature extraction without increasing model complexity. The multi-level ELBP branch optimizes texture feature selection, ensuring a balance between performance and efficiency. Extensive ablation studies highlight the importance of multi-level texture representations in appropriate color spaces for extracting salient spoof detection cues. Experimental results confirm its effectiveness, demonstrating superior intra-dataset performance and strong generalization compared to other state-of-the-art methods.

2. The second proposed solution, the MF^{2}ShrT framework, advances multi-

modal feature fusion by leveraging overlapping patches and shared-layered Vision Transformers (ViTs) to enhance local contextual representations. The SharLViT mechanism optimizes feature learning while reducing computational complexity. Additionally, a T-Encoder-based Hybrid Feature Block effectively captures inter-modal dependencies, and an adaptive fusion mechanism dynamically weights RGB, Hybrid, and RID modalities to emphasize salient information. Extensive evaluations on the CASIA-SURF and WMCA datasets demonstrate the framework's competitive performance, achieving an optimal balance between accuracy and efficiency.

3. The third proposed solution, the Bi-STAM framework, addresses generalization challenges by effectively capturing dynamic motion patterns through bi-directional temporal differences. The Temporal Adaptive Block (TAB) balances static and dynamic features, while the Spatial-Texture Adaptive Block (SAB) refines critical texture cues essential for spoof detection. The fused features significantly enhance classification accuracy, demonstrating resilience against various spoofing attacks across diverse conditions.

4. The fourth proposed solution, PolarSentinelGAN, enhances generalization against diverse and unforeseen attacks by leveraging dual polarized attention (DPAttn) with RGB and MSRCP representations. This framework generates depth maps to facilitate robust live/spoof classification. A Feed Forward Block (FFB) efficiently propagates DPAttn-guided features within the generator, while latent variables further enhance spoof discrimination, improving the model's ability to generalize across scenarios. Experimental results demonstrate state-of-the-art performance across multiple benchmark datasets, supported by a comprehensive ablation study that underscores its effectiveness in addressing key challenges in face anti-spoofing.

7.2 Resource Efficiency of the Proposed FAS Frameworks

To assess the practicality of the proposed face anti-spoofing frameworks, it

Table 7.1 Comparison of Proposed Frameworks: Resource Usage and Practical Applicability

Methodology	Core Strength	Computational Load	Throughput/ Inference Performance	Primary Trade-offs
Two-Stream ELBP + Modified Xception	Lightweight texture + deep multi-level features	Low–Moderate; SE-enhanced backbone without parameter growth	~0.82s per 32-image batch (~40 FPS)	Limited modality fusion; primarily static features
MF^{2}ShrT	Local context via overlapping patches + parameter sharing	Moderate; optimized by shared ViT	~125 multimodal samples/s (≈4–5× standard camera rate)	Limited temporal modeling; relies on static cues
Bi-STAM	Bidirectional motion + adaptive spatio-temporal attention	Moderate–High; temporal + attention blocks	Dataset-level real-time; best suited to batch video	Higher memory footprint due to temporal modeling
PolarSentinelGAN	Depth generation + multi-modal self-attention	High; generative inference + feature fusion	Real-time feasible on dedicated GPU environments	Higher inference cost; requires GPU or hybrid edge-cloud

is essential to examine not only their classification performance but also their computational demands, inference efficiency, and suitability for deployment in real-world environments. Beyond architectural sophistication and accuracy, operational usability depends on maintaining low latency, handling variable input conditions, and functioning effectively under hardware constraints common to consumer and enterprise environments. Accordingly, a comparative evaluation of resource utilization, throughput, and deployment feasibility is provided to illustrate how each framework meets the requirements of practical FAS applications. Table 7.1 summarizes these aspects by relating the core strengths of each method to its computational demands, inference characteristics, and deployment potential.

Collectively, the four frameworks represent a progressive spectrum of design philosophies—from lightweight feature extractors suitable for embedded or edge-based deployments, to temporal and generative architectures optimized for robustness in dynamic or adversarial settings. The two-stream method integrates well with constrained hardware, the multimodal shared-layer ViT enables real-time performance, the temporal modeling approach enhances generalization in high-security environments, and the GAN-based solution delivers strong cross-domain resilience. This holistic perspective underscores that model selection should be guided by deployment context and resource availability, rather than performance alone, ensuring that the chosen system aligns with intended operational environments.

7.3 Future Scope/ Directions and Social Impact

Building on the current advancements, several promising research directions can further strengthen the reliability, generalization, and real-world applicability of face anti-spoofing systems:

- Develop advanced data augmentation strategies using adversarial learning, GANs, and cross-domain synthesis to generate realistic spoof samples.
- Establish realistic, domain-aware open-set evaluation protocols aligned with operational environments.
- Strengthen texture-based spoof detection using robust descriptors and hybrid deep–handcrafted features.
- Leverage transfer learning and disentangled representation learning to isolate spoof cues from nuisance factors.
- Prioritize temporal cues such as facial motion and rPPG for dynamic, context-aware spoof detection.
- Design lightweight, efficient architectures using pruning, quantization, and transformers for real-time deployment.
- Enhance unseen attack detection through one-class learning, anomaly detection, and few-/zero-shot methods.
- Utilize self-supervised and semi-supervised learning to extract richer representations from unlabeled data.
- Incorporate federated learning for privacy-preserving collaborative training

across distributed devices.

- Advance explainable AI (XAI) techniques to improve transparency, fairness, and interpretability.
- Strengthen adversarial robustness to defend against emerging digital and physical attack vectors.
- Explore meta-learning and reinforcement learning for adaptive and generalizable anti-spoofing models.

Social Impact: Developing a deep learning framework for face anti-spoofing carries substantial social benefits, strengthening the security of biometric authentication systems across diverse sectors. This technology effectively prevents identity fraud in banking, surveillance, and digital access control, reducing financial losses and strengthening privacy protection. By improving the reliability of facial recognition, it fosters greater trust in automated systems, encouraging wider adoption in everyday applications. Advanced anti-spoofing solutions contribute to ethical AI deployment by ensuring fairness and generalization across diverse populations, promoting inclusivity, and reducing potential biases. The economic benefits include decreased fraud-related losses and stimulated growth in industries relying on secure biometric authentication. As this technology evolves, it sets new standards for responsible AI use, balancing enhanced security with respect for individual privacy rights. Ultimately, it plays a crucial role in shaping a more secure, inclusive, and trustworthy digital future for society.

REFERENCES

[1] N. K. Ratha, J. H. Connell and R. M. Bol, “An Analysis of Minutiae Matching Strength,” in *3rd International Conference on Audio- and Video-Based Biometric Person Authentication*, Halmstad, Sweden, 2001.

[2] J. Hu, L. Shen and G. Sun, “Squeeze-and-Excitation Networks,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018 .

[3] S. Woo, J. Park, J.-Y. Lee and I. S. Kweon, “CBAM: Convolutional Block Attention Module,” *arXiv:1807.06521 [cs.CV],* 2018.

[4] J. Galbally, S. Marcel and J. Fierrez, “Biometric Antispoofing Methods: A Survey in Face Recognition,” *IEEE Access,* vol. 2, pp. 1530 - 1552, 2014.

[5] Z. Rui and Z. Yan, “A Survey on Biometric Authentication: Toward Secure and Privacy-Preserving Identification,” *IEEE Access,* vol. 7, pp. 5994 - 6009, 2018.

[6] P. Padma and S. Selvaraj, “A Survey on Biometric based Authentication in Cloud Computing,” in *International Conference on Information and Communication Technology (ICICT)*, Coimbatore, India, 2016.

[7] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi and S. Z. Li, “A Face Antispoofing Database with Diverse Attacks,” in *5th IAPR International Conference on Biometrics*, New Delhi, India, 2012.

[8] D. Wen, H. Han and A. K. Jain, “Face Spoof Detection with Image Distortion Analysis,” *IEEE Transactions on Information Forensics and Security,* vol. 10, no. 4, p. 746–761, 2015.

[9] K. Delac and M. Grgic, “A Survey of Biometric Recognition Methods,” in *46th International Symposium on Electronics in Marine*, Zadar, 2004.

[10] A. Hadid, “Face Biometrics Under Spoofing Attacks: Vulnerabilities, Countermeasures, Open Issues, and Research Directions,” in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Columbus, 2014.

[11] E. A. Raheem, S. M. S. Ahmad and w. A. w. adnan, “Insight on Face Liveness Detection: A Systematic Literature Review,” *International Journal of Electrical and Computer Engineering,* vol. 9, no. 6, p. 5165~5175, 2019.

[12] R. Tolosana, R. V. Rodriguez, J. Fierrez, A. Morales and J. O. Garcia, “DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection,” *Information Fusion,* vol. 64, pp. 131-148, 2020.

[13] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei and G. Zhao, “Deep Learning for Face Anti-Spoofing: A Survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 45, no. 5, pp. 5609-5631, 2023.

[14] S. Jia, X. Li, C. Hu, G. Guo and Z. Xu, “3D Face Anti-spoofing with Factorized Bilinear Coding,” *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 31, no. 10, pp. 4031-4045, 2020.

[15] A. Liu, C. Zhao, Z. Yu, J. Wan, A. Su, X. Liu, Z. Tan, S. Escalera, J. Xing, Y. Liang, G. Guo, Z. Lei, S. Z. Li, and D. Zhang, “Contrastive Context-Aware Learning for 3D High-Fidelity Mask Face Presentation Attack Detection,” *IEEE Transactions on Information Forensics and Security,* vol. 17, pp. 2497 - 2507, 2021.

[16] G. Heusch, A. George, D. Geissbuhler, Z. Mostaani and S. Marcel, “Deep Models and Shortwave Infrared Information to Detect Face Presentation Attacks,” *IEEE*

Transactions on Biometrics, Behavior, and Identity Science, vol. 2, no. 4, 2020.

[17] G. Pan, L. Sun, Z. Wu and S. Lao, "Eyeblink-based Anti-Spoofing in Face Recognition from a Generic Webcamera," in *11th IEEE International Conference on Computer Vision*, Shenzhen, China, 2007.

[18] D. F. Smith, A. Wiliem and B. C. Lovell, "Face Recognition on Consumer Devices: Reflections on Replay Attacks," *IEEE Transactions on Information Forensics and Security,* vol. 10, no. 4, pp. 736 - 745, 2015.

[19] P. A. Johnson, B. Tan and S. Schuckers, "Multimodal Fusion Vulnerability to Non-Zero Effort (spoof) Imposters," in *IEEE International Workshop on Information Forensics and Security*, Seattle, 2010.

[20] "IBG,"Biometrics Market and Industry Report 2009-2014"," International Biometrics Group, Virginia, USA, 2008.

[21] B. Gipp, J. Beel and I. Rössling, ePassport: The World's New Electronic Passport., scott valley,CA, USA: CreateSpace, 2007.

[22] Z. Boulkenafet, J. Komulainen and A. Hadid, "Face Anti-Spoofing based on Color Texture Analysis," in *IEEE International Conference on Image Processing (ICIP)*, Canada, 2015.

[23] A. Pinto, W. R. Schwartz, H. Pedrini and A. d. R. Rocha, "Using Visual Rhythms for Detecting Video-based Facial Spoof Attacks," *IEEE Transactions on Information Forensics and Security,* vol. 10, no. 5, pp. 1025 - 1038, 2015.

[24] Z. Boulkenafet, J. Komulainen and A. Hadid, "Face Spoofing Detection Using Color Texture Analysis," *IEEE Transactions on Information Forensics and Security,* vol. 11, no. 8, pp. 1818 - 1830, 2016.

[25] R. J. Raghavendra and R. S. Kunte, "A Novel Feature Descriptor for Face Anti-Spoofing Using Texture Based Method," *Cybernetics and Information Technologies,* vol. 20, no. 3, p. 159–176, 2020.

[26] Z. Boulkenafet, J. Komulainen and A. Hadid, "Face Antispoofing using Speeded-up Robust Features and Fisher Vector Encoding," *IEEE Signal Processing Letters,* vol. 24, no. 2, pp. 141 - 145, 2017.

[27] D. Das and S. Chakraborty, "Face Liveness Detection based on Frequency and Micro-Texture Analysis," in *International Conference on Advances in Engineering & Technology Research (ICAETR)*, Unnao, 2014.

[28] Z. Akhtar and G. L. Foresti, "Face Spoof Attack Recognition Using Discriminative Image Patches," *Journal of Electrical and Computer Engineering,* no. 1, pp. 1-14, 2016.

[29] A. Pinto, H. Pedrini, W. R. Schwartz and A. Rocha, "Face Spoofing Detection through Visual Codebooks of Spectral Temporal Cubes," *IEEE Transactions on Image Processing,* vol. 24, no. 12, pp. 4726 - 4740, 2015.

[30] X. Tan, Y. Li, J. Liu and L. Jiang, "Face Liveness Detection from a Single Image with Sparse Low Rank Bilinear Discriminative Model," in *Computer Vision– European Conference on Computer Vision*, 2010.

[31] J. Li, Y. Wang, T. Tan and A. K. Jain, "Live Face Detection Based on the Analysis of Fourier Spectra," in *Proceedings of SPIE - The International Society for Optical Engineering*, 2004.

[32] K. Kollreider, H. Fronthaler and J. Bigun, "Verifying Liveness by Multiple Experts in Face Biometrics," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Anchorage, AK, USA, 2008.

[33] M. Singh and A. S. Arora, “A Robust Anti-spoofing Technique for Face Liveliness Detection with Morphological Operations,” *Optik,* vol. 139, pp. 347-354, 2017.

[34] K. Kollreider, H. Fronthaler, M. I. Faraj and J. Bigun, “Real-Time Face Detection and Motion Analysis With Application in “Liveness” Assessment,” *IEEE Transactions on Information Forensics and Security,* vol. 2, no. 3, pp. 548 - 558, 2007.

[35] K. Kollreider, H. Fronthaler and J. Bigun, “Non-intrusive Liveness Detection by Face Images,” *Image and Vision Computing,* vol. 27, no. 3, pp. 233-244, 2009.

[36] W. Yin, Y. Ming and L. Tian, “A Face Anti-Spoofing Method Based on Optical Flow Field,” in *IEEE 13th International Conference on Signal Processing (ICSP)*, Chengdu, China, 2016.

[37] S. Liu, P. C. Yuen, S. Zhang and G. Zhao, “3D Mask Face Anti-spoofing with Remote Photoplethysmography,” in *Computer Vision – European Conference on Computer Vision*, Amsterdam, 2016.

[38] L. Feng, L.-M. Po, Y. Li, X. Xu, F. Yuan, T. C.-H. Cheung and K.-W. Cheung, “Integration of Image Quality and Motion Cues for Face Anti-spoofing: A Neural Network Approach,” *Journal of Visual Communication and Image Representation,* vol. 38, pp. 451-460, 2016.

[39] W. R. Schwartz , A. Rocha and H. Pedrini, “Face Spoofing Detection throughPartial Least Squares and Low-Level Descriptors,” in *International Joint Conference on Biometrics (IJCB)*, Washington DC, United States, 2011.

[40] X. Li, J. Komulainen, G. Zhao, P.-C. Yuen and M. Pietikainen, “Generalized Face Anti-spoofing by Detecting Pulse from Face Videos,” in *23rd International Conference on Pattern Recognition (ICPR)*, Cancún, México, 2016.

[41] Z. Boulkenafet, J. Komulainen, X. Feng and A. Hadid, “Scale Space Texture Analysis for Face Anti-spoofing,” in *International Conference on Biometrics (ICB)*, Halmstad, 2016.

[42] A. Antil and C. Dhiman, “Two Stream RGB-LBP Based Transfer Learning Model for Face Anti-spoofing,” in *International Conference on Computer Vision and Image Processing*, India, 2023.

[43] G. Huang, Z. Liu, L. V. D. Maaten and K. Q. Weinberger, “Densely Connected Convolutional Networks,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017.

[44] S. Loffe and C. Szegedy, “Batch normalization: Accelerating deep Network Training by Reducing Internal Covariate Shift,” in *32nd International Conference on Machine Learning*, Lille, France, 2015.

[45] X. Shu , X. Li, X. Zuo, D. Xu and J. Shi, “Face Spoofing Detection based on Multi-Scale Color Inversion Dual-stream Convolutional Neural Network,” *Expert Systems With Applications,* vol. 224, 2023.

[46] J. Yang, Z. Lei and S. Z. Li, “Learn Convolutional Neural Network for Face Anti-Spoofing,” *arXiv:1408.5601 [cs.CV],* 2014.

[47] O. Lucena, A. Junior, V. Moia, R. Souza, E. Valle and R. Lotufo, “Transfer Learning Using Convolutional NeuralNetworks for Face Anti-Spoofing,” in *International Conference Image Analysis and Recognition ((ICIAR)*, Montreal, Canada, 2017.

[48] Y. A. U. Rehman, L. M. Po and M. Liu, “Livenet: Improving Features Generalization for Face Liveness Detection using Convolution Neural Networks,” *Expert Systems with Applications,* vol. 108, pp. 159-169, 2018.

[49] X. Yang, W. Luo, L. Bao, Y. Gao, D. Gong, S. Zheng, Z. Li and W. Liu, “Face Anti-

Spoofing: Model Matters, So Does Data," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019.

[50] U. Muhammad, T. Holmberg, W. C. d. Melo and A. Hadid, "Face Anti-Spoofing via Sample Learning based Recurrent Neural Network," in *British Machine Vision Conference (BMVC)*, Cardiff, UK, 2019.

[51] A. Pinto, S. Goldenstein, A. Ferreira, T. Carvalho, H. Pedrini and A. Rocha, "Leveraging Shape, Reflectance and Albedo From Shading for Face Presentation Attack Detection," *IEEE Transactions on Information Forensics and Security,* vol. 15, pp. 3347 - 3358, 2020.

[52] G. B. d. Souza, J. P. Papa and A. N. Marana, "On the Learning of Deep Local Features for Robust Face Spoofing Detection," in *31st SIBGRAPI Conference on Graphics, Patterns and Images*, Parana, Brazil, 2018.

[53] H. Chen, G. Hu, Z. Lei, Y. Chen, N. M. Robertson and S. Z. Li, "Attention-Based Two-Stream Convolutional Networks for Face Spoofing Detection," *IEEE Transactions on Information Forensics and Security,* vol. 15, pp. 578 - 593, 2020.

[54] D. Deb and A. K. Jain, "Look Locally Infer Globally: A Generalizable Face Anti-Spoofing Approach," *IEEE Transactions on Information Forensics and Security,* vol. 16, p. 1143–1157, 2020.

[55] A. George and S. Marcel, "On the Effectiveness of Vision Transformers for Zero-shot Face Anti-Spoofing," in *IEEE International Joint Conference on Biometrics (IJCB)*, Shenzhen, China, 2021.

[56] R. Cai, H. Li , S. Wang, C. Chen and A. C. Kot, "DRL-FAS: A Novel Framework Based on Deep Reinforcement Learning for Face Anti-Spoofing," *IEEE Transactions on Information Forensics and Security,* vol. 16, pp. 937 - 951, 2021.

[57] Y. Kong , X. Li , G. Hao and C. Liu , "Face Anti-Spoofing Method Based on Residual Network with Channel Attention Mechanism," *Electronics,* vol. 11, no. 19, p. 3056, 2022.

[58] C. Lin, Z. Liao, P. Zhou, J. Hu and B. Ni, "Live Face Verification with Multiple Instantialized Local Homographic Parameterization," in *27th International Joint Conference on Artificial Intelligence (IJCAI)*, Stockholm,Sweden, 2018.

[59] H. Hao, M. Pei and M. Zhao, "Face Liveness Detection Based on Client Identity Using Siamese Network," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, Xi'an, China, 2019.

[60] Y. A. U. Rehman, L.-M. Po, M. Liu, Z. Zou, W. Ou and Y. Zhao, "Face Liveness Detection using Convolutional-Features Fusion of Real and Deep Network Generated Face Images," *Journal of Visual Communication and Image Representation,* vol. 59, pp. 574-582, 2019.

[61] L. Li, Z. Xia, X. Jiang, F. Roli and X. Feng, "Compactnet: Learning a Compact Space for Face Presentation Attack Detection," *Neurocomputing,* vol. 409, pp. 191-207, 2020.

[62] X. Xu, Y. Xiong and W. Xia, " On Improving Temporal Consistency for Online Face Liveness Detection System," in *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Montreal, BC, Canada, 2021.

[63] B. Chen, W. Yang, H. Li, S. Wang and S. Kwong, "Camera Invariant Feature Learning for Generalized Face Anti-Spoofing," *IEEE Transactions on Information Forensics and Security,* vol. 16, pp. 2477 - 2492, 2021.

[64] W. R. Almeida, F. A. Andaló, R. Padilha, G. Bertocco, W. Dias, R. d. S. Torres, J. Wainer and A. Rocha, "Detecting Face Presentation Attacks in Mobile Devices with a

Patch-based CNN and a Sensor-Aware Loss Function," *PLoS One,* vol. 15, no. 9, 2020.

[65] C.-Y. Wang, Y.-D. Lu, S.-T. Yang and S.-H. Lai, "PatchNet: A Simple Face Anti-Spoofing Framework via Fine-Grained Patch Recognition," *arXiv:2203.14325 [cs.CV],* 2022.

[66] Y.-H. Huang, J.-W. Hsieh, M.-C. Chang, L. Ke, S. Lyu and A. S. Santra, "Multi-Teacher Single-Student Visual Transformer with Multi-Level Attention for Face Spoofing Detection," in *British Machine Vision Conference*, 2021.

[67] T. Qiao, J. Wu, N. Zheng, M. Xu and X. Luo, "FGDNet: Fine-Grained Detection Network Towards Face Anti-Spoofing," *IEEE Transactions on Multimedia,* vol. 25, pp. 7350-736, 2023.

[68] T. Wang, J. Yang, Z. Lei, S. Liao and S. Z. Li, "Face Liveness Detection using 3D Structure Recovered from a Single Camera," in *International Conference on Biometrics (ICB)*, Madrid, Spain, 2013.

[69] X. Li, J. Wan, Y. Jin, A. Liu, G. Guo and S. Z. Li, "3DPC-Net: 3D Point Cloud Network for Face Anti-spoofing," in *IJCB*, Houston, TX, USA, 2020.

[70] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou and G. Zhao, "Searching Central Difference Convolutional Networks for Face Anti-Spoofing," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020.

[71] W. Zheng, M. Yue, S. Zhao and S. Liu, "Attention-Based Spatial-Temporal Multi-Scale Network for Face Anti-Spoofing," *IEEE Transactions on Biometrics, Behavior, and Identity Science,* vol. 3, no. 3, pp. 296 - 307, 2021.

[72] L. Zhang, N. Sun, X. Wu and D. Luo, "Advanced Face Anti-Spoofing with Depth Segmentation," in *International Joint Conference on Neural Networks (IJCNN)*, Padua, Italy, 2022.

[73] Y. Atoum, Y. Liu, A. Jourabloo and X. Liu, "Face Anti-Spoofing using Patch and Depth-based CNNs," in *IEEE International Joint Conference on Biometrics (IJCB)*, Denver, CO, USA, 2017.

[74] D. Peng, J. Xiao, R. Zhu and G. Gao, "Ts-Fen: Probing Feature Selection Strategy for Face Anti-Spoofing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020.

[75] T. Kim, Y. Kim, I. Kim and D. Kim, "BASN: Enriching Feature Representation Using Bipartite Auxiliary Supervisions for Face Anti-Spoofing," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul,Korea, 2019.

[76] Y. Zhang, Z. Yin, Y. Li, G. Yin, J. Yan, J. Shao and Z. Liu, "CelebA-Spoof: Large-Scale Face Anti-Spoofing Dataset with Rich Annotations," in *European Conference on Computer Vision*, Glasgow, United Kingdom, 2020.

[77] M. Fang, N. Damer, F. Kirchbuchner and A. Kuijper, "Learnable Multi-level Frequency Decomposition and Hierarchical Attention Mechanism for Generalized Face Presentation Attack Detection," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 2022.

[78] Y. Liu, J. Stehouwer, A. Jourabloo and X. Liu, "Deep Tree Learning for Zero-Shot Face Anti-Spoofing," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, California, 2019.

[79] Z. Yu, X. Li, J. Shi, Z. Xia and G. Zhao, "Revisiting Pixel-Wise Supervision for Face Anti-Spoofing," *IEEE Transactions on Biometrics, Behavior and Identity Science,* vol. 3, no. 3, pp. 285-295, 2021.

[80] Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li and G. Zhao, “NAS-FAS: Static-Dynamic Central Difference Network Search for Face Anti-Spoofing,” *IEEE Transactions on Pattern Analysis and Machine Intelligence ,* vol. 43, no. 9, pp. 3005 - 3023, 2021.

[81] Z. Wang, Z. Yu, C. Zhao and X. Zhu, “Deep Spatial Gradient and Temporal Depth Learning for Face Anti-Spoofing,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020.

[82] Z. Yu, Y. Qin, H. Zhao, X. Li and G. Zhao, “Dual-Cross Central Difference Network for Face Anti-Spoofing,” in *30th International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[83] H. Wu, D. Zeng, Y. Hu, H. Shi and T. Mei, “Dual Spoof Disentanglement Generation for Face Anti-spoofing with Depth Uncertainty Learning,” *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 32, no. 7, pp. 4626 - 4638, 2022.

[84] Z. Wang, Y. Xu, L. Wu, H. Han, Y. Ma and G. Ma, “Multi-Perspective Features Learning for Face Anti-Spoofing,” in *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Montreal, BC, Canada, 2021.

[85] C. Wang and J. Zhou, “An Adaptive Index Smoothing Loss for Face Anti-Spoofing,” *Pattern Recognition Letters,* vol. 153, no. 2, pp. 168-175, 2022.

[86] Z. Wang, Q. Wang, W. Deng and G. Guo, “Learning Multi-Granularity Temporal Characteristics for Face Anti-Spoofing,” *IEEE Transactions on Information Forensics and Security,* vol. 17, pp. 1254 - 1269, 2022.

[87] Z. Wang, Q. Wang, W. Deng and G. Guo, “Face Anti-Spoofing Using Transformers With Relation-Aware Mechanism,” *IEEE Transactions on Biometrics, Behavior, and Identity Science ,* vol. 4, no. 3, pp. 439 - 450, 2022.

[88] Y. Ma, L. Wu, Z. Lia and F. liu, “A Novel Face Presentation Attack Detection Scheme based on Multi-Regional Convolutional Neural Networks,” *Pattern Recognition Letters,* vol. 131, pp. 261-267, 2020.

[89] A. George and S. Marcel, “Deep Pixel-wise Binary Supervision for Face Presentation Attack Detection,” in *International Conference on Biometrics (ICB)*, Crete, Greece, 2019.

[90] M. S. Hossain, L. Rupty, K. Roy and M. Hasan, “A-DeepPixBis: Attentional Angular Margin for Face Anti-Spoofing,” in *Digital Image Computing: Techniques and Applications (DICTA)*, Melbourne, Australia, 2020.

[91] Z. Yu, Y. Qin, X. Xu, C. Zhao, Z. Wang, Z. Lei and G. Zhao, “Auto-Fas: Searching Lightweight Networks for Face Anti-Spoofing,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020.

[92] Z. Yu, X. Li , X. Niu, J. Shi and G. Zhao , “Face Anti-Spoofing with Human Material Perception,” in *European Conference on Computer Vision*, Glasgow, 2020.

[93] Y. Liu, A. Jourabloo and X. Liu, “Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.

[94] C. Hu, J. Cao, K.-Y. Zhang, T. Yao, s. Ding and L. Ma, “Structure Destruction and Content Combination for Generalizable Anti-Spoofing,” *IEEE Transactions on Biometrics, Behavior, and Identity Science,* vol. 4, no. 4 , pp. 508 - 521, 2022.

[95] K. Roy , M. Hasan , L. Rupty , M. S. Hossain , S. Sengupta , S. N. Taus and N. Mohammed , “Bi-FPNFAS: Bi-Directional Feature Pyramid Network for Pixel-Wise Face Anti-Spoofing by Leveraging Fourier Spectra,” *Sensors,* vol. 21, no. 8, p. 2799, 2021.

[96] W. Sun, Y. Song, C. Chen, J. Huang and A. C. Kot, "Face Spoofing Detection Based on Local Ternary Label Supervision in Fully Convolutional Networks," *IEEE Transactions on Information Forensics and Security,* vol. 15, pp. 3181 - 3196, 2020.

[97] M. Hasan, K. Roy, L. Rupty, M. S. Hossain, S. Sengupta, S. N. Taus and N. Mohammed, "MHASAN: Multi-Head Angular Self Attention Network for Spoof Detection," in *26th International Conference on Pattern Recognition (ICPR)*, Montreal, QC, Canada, 2022.

[98] Y. Bian, P. Zhang, J. Wang, C. Wang and S. Pu, "Learning Multiple Explainable and Generalizable Cues for Face Anti-Spoofing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, 2022.

[99] Y. Liu and X. Liu, "Physics-Guided Spoof Trace Disentanglement for Generic Face Anti-Spoofing," *arXiv:2012.05185 [cs.CV],* 2020.

[100] Y. Liu and X. Liu, "Spoof Trace Disentanglement for Generic Face Anti-Spoofing," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 45, no. 3, pp. 3813 - 3830, 2023.

[101] B. Zhang, X. Zhu, X. Zhang and Z. Lei, "Modeling Spoof Noise by De-spoofing Diffusion and its Application in Face Anti-spoofing," in *IEEE International Joint Conference on Biometrics (IJCB)*, Ljubljana, Slovenia, 2023.

[102] A. Jourabloo, Y. Liu and X. Liu, "Face De-spoofing: Anti-spoofing via Noise Modeling," in *European Conference on Computer Vision*, Munich, Germany, 2018.

[103] M. O. Alassafi, M. S. Ibrahim, I. Naseem, R. AlGhamdi, F. A. Kateb, H. M. Oqaibi, A. A. Alshdadi and S. A. Yusuf, "A Novel Deep Learning Architecture With Image Diffusion for Robust Face Presentation Attack Detection," *IEEE Access,* vol. 11, pp. 59204 - 59216, 2023.

[104] Y. Wang, X. Song, T. Xu, Z. Feng and X.-J. Wu, "From RGB to Depth: Domain Transfer Network for Face Anti-Spoofing," *IEEE Transactions on Information Forensics and Security,* vol. 16, p. 4280–4290, 2021.

[105] J. Stehouwer, A. Jourabloo, Y. Liu and X. Liu, "Noise Modeling, Synthesis and Classification for Generic Object Anti-Spoofing," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020.

[106] Y. Liu, J. Stehouwer and X. Liu, "On Disentangling Spoof Trace for Generic Face Anti-spoofing," in *European Conference on Computer Vision*, Glasgow, 2020.

[107] K.-Y. Zhang, T. Yao, J. Zhang, Y. Tai, S. Ding, J. Li, F. Huang, H. Song and L. Ma, "Face Anti-Spoofing via Disentangled Representation Learning," in *European Conference on Computer Vision*, Glasgow, 2020.

[108] Y. Qin, Z. Yu, L. Yan, Z. Wang, C. Zhao and Z. Lei, "Meta-Teacher For Face Anti-Spoofing," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 44, no. 10, 2021.

[109] Y.-C. Wang, C.-Y. Wang and S.-H. Lai, "Disentangled Representation with Dual-stage Feature Learning for Face Anti-spoofing," in *IEEE/CVF WACV*, France, 2022.

[110] Z. Zhang, H. Cheng, W. Li and P. Wang, "An Improved GAN-based Depth Estimation Network for Face Anti-Spoofing," in *9th International Conference on Computing and Artificial Intelligence (ICCV)*, Tianjin, 2023.

[111] H. Li, W. Li, H. Cao, S. Wang, F. Huang and A. C. Kot, "Unsupervised Domain Adaptation for Face Anti-Spoofing," *IEEE Transactions on Information Forensics and Security,* vol. 13, no. 7, p. 1794–1809, 2018.

[112] G. Wang, H. Han, S. Shan and X. Chen, "Improving Cross-database Face Presentation

Attack Detection via Adversarial Domain Adaptation," in *International Conference on Biometrics (ICB)*, Crete, Greece, 2019.

[113] G. Wang, H. Han, S. Shan and X. Chen, "Unsupervised Adversarial Domain Adaptation for Cross-Domain Face Presentation Attack Detection," *IEEE Transactions on Information Forensics and Security,* vol. 16, pp. 56-69, 2020.

[114] Y. S. El-Din, M. N. Moustafa and H. Mahdi, "Deep Convolutional Neural Networks for Face and Iris Presentation Attack Detection: Survey and Case Study," *IET Biometrics,* vol. 9, no. 5, pp. 179-193, 2020.

[115] A. Panwar, P. Singh, S. Saha, D. P. Paudel and L. V. Gool, "Unsupervised Compound Domain Adaptation for Face Anti-Spoofing," in *16th IEEE International Conference on Automatic Face and Gesture Recognition* , Jodhpur, India, 2021.

[116] Y. Jia, J. Zhang, S. Shan and X. Chen, "Unified Unsupervised and Semi-Supervised Domain Adaptation Network for Cross-Scenario Face Anti-spoofing," *Pattern Recognition,* vol. 115, p. 107888, 2021.

[117] X. Tu, Z. Ma, J. Zhao, G. Du, M. Xie and J. Feng, "Learning Generalizable and Identity-Discriminative Representations for Face Anti-Spoofing," *ACM Transactions on Intelligent Systems and Technology,* vol. 11, no. 5, pp. 1-19, 2019.

[118] A. Mohammadi, S. Bhattacharjee and S. Marcel, "Improving Cross-Dataset Performance of Face Presentation Attack Detection Systems Using Face Recognition Datasets," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020.

[119] A. Mohammadi, S. Bhattacharjee and S. Marcel, "Domain Adaptation for Generalization of Face Presentation Attack Detection in Mobile Settengs with Minimal Information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020.

[120] H. Li, S. Wang, P. He and A. Roch, "Face Anti-Spoofing With Deep Neural Network Distillation," *IEEE Journal of Selected Topics in Signal Processing,* vol. 14, no. 5, pp. 933 - 946, 2020.

[121] Y. Qin, C. Zhao, X. Zhu, Z. Wang, Z. Yu, T. Fu, F. Zhou, J. Shi and Z. Lei, "Learning Meta Model for Zero-and Few-shot Face Anti-spoofing," in *AAAI Conference on Artificial Intelligence*, New York, USA, 2020.

[122] D. Pérez-Cabo, D. Jiménez-Cabello, A. Costa-Pazo and R. J. López-Sastre, "Learning to Learn Face-PAD: A Lifelong Learning Approach," in *IEEE International Joint Conference on Biometrics (IJCB)*, Houston, TX, USA, 2020.

[123] D. Perez-Cabo, D. Jimenez-Cabello, A. Costa-Pazo and R. J. Lopez-Sastre, "Deep Anomaly Detection for Generalized Face Anti-Spoofing," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, 2019.

[124] S. Fatemifar, S. R. Arashloo, M. Awais and J. Kittler, "Spoofing Attack Detection by Anomaly Detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK, 2019.

[125] S. R. Arashloo, J. Kittler and W. Christmas, "An Anomaly Detection Approach to Face Spoofing Detection: A New Formulation and Evaluation Protocol," *IEEE Access,* vol. 5, pp. 13868 - 13882, 2017.

[126] O. Nikisins, A. Mohammadi, A. Anjos and S. Marcel, "On Effectiveness of Anomaly Detection Approaches against Unseen Presentation Attacks in Face Anti-spoofing," in *International Conference on Biometrics (ICB)*, Gold Coast, QLD, Australia, 2018.

[127] A. George and S. Marcel, "Learning One Class Representations for Face Presentation

Attack Detection Using Multi-Channel Convolutional Neural Networks," *IEEE Transactions on Information Forensics and Security,* vol. 16, pp. 361 - 375, 2020.

[128] S. Fatemifar, M. Awais, A. Akbari and J. Kittler, "A Stacking Ensemble for Anomaly Based Client-Specific Face Spoofing Detection," in *IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates, 2020.

[129] Z. Li, H. Li, K.-Y. Lam and A. C. Kot, "Unseen Face Presentation Attack Detection with Hypersphere Loss," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020.

[130] Y. Baweja, P. Oza, P. Perera and V. M. Patel, "Anomaly Detection-Based Unknown Face Presentation Attack Detection," in *IEEE International Joint Conference on Biometrics (IJCB)*, TX,USA, 2020.

[131] X. Dong, H. Liu, W. Cai, P. Lv and Z. Yu, "Open Set Face Anti-Spoofing in Unseen Attacks," in *29th ACM International Conference on Multimedia*, Chengdu, China, 2021.

[132] A. Li, Z. Tan, X. Li, J. Wan, S. Escalera, G. Guo and S. Z. Li, "CASIA-SURF CeFA: A Benchmark for Multi-modal Cross-ethnicity Face Anti-spoofing," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 2021.

[133] H. Kuang, R. Ji, H. Liu, S. Zhang, X. Sun, F. Huang and B. Zhang, "Multi-modal Multi-layer Fusion Network with Average Binary Center Loss for Face Anti-spoofing," in *27th ACM International Conference on Multimedia*, Nice, France, 2019.

[134] A. Parkin and O. Grinchuk, "Recognizing Multi-Modal Face Spoofing With Face Recognition Networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, 2019.

[135] T. Shen, Y. Huang and Z. Tong, "FaceBagNet: Bag-of-local-features Model for Multi-modal Face Anti-spoofing," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, 2019.

[136] S. Zhang, X. Wang, A. Liu, C. Zhao, J. Wan, S. Escalera, H. Shi, Z. Wang and S. Z. Li, "A Dataset and Benchmark for Large-Scale Multi-Modal Face Anti-Spoofing," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019.

[137] Z. Yu, Y. Qin, X. Li, Z. Wang, C. Zhao, Z. Lei and G. Zhao, "Multi-Modal Face Anti-Spoofing Based on Central Difference Networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA, 2020.

[138] Z. Yu, R. Cai, Y. Cui, X. Liu, Y. Hu and A. Kot, "Rethinking Vision Transformer and Masked Autoencoder in Multimodal Face Anti-Spoofing," *International Journal of Computer Vision,* vol. 132, pp. 5217 - 5238, 2024.

[139] W. Liu, X. Wei, T. Lei, X. Wang, H. Meng and A. K. Nandi, "Data-Fusion-Based Two-Stage Cascade Framework for Multimodality Face Anti-Spoofing," *IEEE Transactions on Cognitive and Developmental Systems,* vol. 14, no. 2, pp. 672-683, 2021.

[140] A. Antil and C. Dhiman, "MF2ShrT: Multimodal Feature Fusion Using Shared Layered Transformer for Face Anti-spoofing," *ACM Transactions on Multimedia Computing, Communications, and Applications,* vol. 20, no. 6, pp. 1-21, 2024.

[141] A. Liu, Z. Tan, J. Wan, Y. Liang, Z. Lei, G. Guo and S. Z. Li, "Face Anti-Spoofing via Adversarial Cross-Modality Translation," *IEEE Transactions on Information Forensics and Security ,* vol. 16, pp. 2759 - 2772, 2021.

[142] P. Zhang, F. Zou, Z. Wu, N. Dai, S. Mark, M. Fu, J. Zhao and K. Li, "FeatherNets:

Convolutional Neural Networks as Light as Feather for Face Anti-spoofing," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[143] A. George and S. Marcel, "Can Your Face Detector Do Anti-spoofing? Face Presentation Attack Detection with a Multi-Channel Face Detector," *arXiv:2006.16836v2 [cs.CV] ,* 2020.

[144] A. George, D. Geissbuhler and S. Marcel, "A Comprehensive Evaluation on Multi-channel Biometric Face Presentation Attack Detection," *arXiv:2202.10286 [cs.CV],* 2022.

[145] W. Wang, F. Wen, H. Zheng, R. Ying and P. Liu, "Conv-MLP: A Convolution and MLP Mixed Model for Multimodal Face Anti-Spoofing," *IEEE Transactions on Information Forensics and Security,* vol. 17, pp. 2284 - 2297, 2022.

[146] F. Jiang, P. Liu, X. Shao and X. Zhou , "Face Anti-Spoofing with Generated Near-Infrared Images," *Multimedia Tools and Applications,* vol. 79, no. 29, p. 21299–21323, 2020.

[147] K. Mallat and J.-L. Dugelay, "Indirect synthetic attack on thermal face biometric systems via visible-to-thermal spectrum conversion," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA, 2021.

[148] X. Song, X. Zhao, L. Fang and T. Lin, "Discriminative Representation Combinations for Accurate Face Spoofing Detection," *Pattern Recognition,* vol. 85, pp. 220-231, 2019.

[149] A. Agarwal, M. Vatsa and R. Singh, "CHIF: Convoluted Histogram Image Features for Detecting Silicone Mask based Face Presentation Attack," in *IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Tampa, FL, USA, 2019.

[150] L. Li and X. Feng, "Face Anti-spoofing via Deep Local Binary Pattern," in *IEEE International Conference on Image Processing (ICIP)*, Beijing, China, 2017.

[151] H. Chen, Y. Chen, X. Tian and R. Jiang, "A Cascade Face Spoofing Detector Based on Face Anti-Spoofing R-CNN and Improved Retinex LBP," *IEEE Access ,* vol. 7, pp. 170116 - 170133, 2019.

[152] P. K. Das, B. Hu, C. Liu, K. Cui, P. Ranjan and G. Xiong, "A New Approach for Face Anti-Spoofing Using Handcrafted and Deep Network Features," in *IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, Zhengzhou, China, 2019.

[153] R. Cai and C. Chen, "Learning Deep Forest with Multi-Scale Local Binary Pattern Features for Face Anti-Spoofing," *arXiv:1910.03850 [cs.CV],* 2019.

[154] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li and A. Hadid, "An Original Face Anti-Spoofing Approach using Partial Convolutional Neural Network," in *6th International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Oulu, Finland, 2016.

[155] M. Asim, Z. Ming and M. Y. Javed, "CNN based Spatio-Temporal Feature Extraction for Face Anti-Spoofing," in *2nd International Conference on Image, Vision and Computing (ICIVC)*, Chengdu, 2017.

[156] R. Shao, X. Lan and P. C. Yuen, "Joint Discriminative Learning of Deep Dynamic Textures for 3D Mask Face Anti-Spoofing," *IEEE Transactions on Information Forensics and Security,* vol. 14, no. 4, pp. 923-938, 2019.

[157] O. Sharifi, "Score-Level-based Face Anti-Spoofing System Using Handcrafted and

Deep Learned Characteristics," *International Journal of Image Graphics and Signal Processing,* vol. 11, no. 2, pp. 15-20, 2019.

[158] Y. A. U. Rehman, L.-M. Po, M. Liu, Z. Zou and W. Ou, "Perturbing Convolutional Feature Maps with Histogram of Oriented Gradients for Face Liveness Detection," in *International Joint Conference: 12th International Conference on Computational Intelligence in Security for Information Systems (CISIS) and 10th International Conference on EUropean Transnational Education (ICEUTE)*, 2019.

[159] Y. A. U. Rehman, L.-M. Po and J. Komulainen, "Enhancing Deep Discriminative Feature Maps via Perturbation for Face Presentation Attack Detection," *Image and Vision Computing,* vol. 94, p. 103858, 2020.

[160] L. Li, Z. Xia, A. Hadid, X. Jiang, H. Zhang and X. Feng, "Replayed Video Attack Detection Based on Motion Blur Analysis," *TIFS,* vol. 14, no. 9, pp. 2246-2261, 2019.

[161] H. Qi, C. Wu, Y. Shi, X. Qi, K. Duan and X. Wang, "A Real-Time Face Detection Method Based on Blink Detection," *IEEE Access,* vol. 11, pp. 28180 - 28189, 2023.

[162] L. Li, Z. Yao, S. Gao, H. Han and Z. Xia, "Face Anti-Spoofing via Jointly Modeling Local Texture and Constructed Depth," *Engineering Applications of Artificial Intelligence,* vol. 133, p. 108345, 2024.

[163] Y. Chen, T. Wang, J. Wang, P. Shi, G. Shan and H. Snoussi, "Towards Good Practices in Face Anti-Spoofing: An Image Reconstruction Based Method," in *Chinese Automation Congress (CAC)*, Hangzhou, China, 2019.

[164] H. Feng, Z. Hong, H. Yue, Y. Chen, K. Wang, J. Han, J. Liu and E. Ding, "Learning Generalized Spoof Cues for Face Anti-spoofing," in *2020*, arXiv:2005.03922 [cs.CV].

[165] C.-H. Liao, W.-C. Chen, H.-T. Liu, Y.-R. Yeh, M.-C. Hu and C.-S. Chen, "Domain Invariant Vision Transformer Learning for Face Anti-spoofing," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, 2023.

[166] A. Khurshid, S. C. Tamayo, E. Fernandes, M. R. Gadelha and M. Teofilo, "A Robust and Real-Time Face Anti-spoofing Method Based on Texture Feature Analysis," in *International Conference on Human-Computer Interaction*, Walt Disney World Swan and Dolphin Resort, Orlando, Florida, USA, 2019.

[167] T. A. Siddiqui, S. Bharadwaj, T. I. Dhamecha, A. Agarwal, M. Vatsa, R. Singh and N. Ratha, "Face Anti-Spoofing with Multifeature Videolet Aggregation," in *23rd International Conference on Pattern Recognition (ICPR)*, Cancun, Mexico, 2017.

[168] S. Liu, X. Lan and P. C. Yuen, "Remote Photoplethysmography Correspondence Feature for 3D Mask Face Presentation Attack Detection," in *European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018.

[169] V. J. and S. , "Detecting Pedestrians usingPpatterns of Motion and Appearance," in *9th IEEE International Conference on Computer Vision*, Nice, France, 2003.

[170] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment using Multitask Cascaded Convolutional Networks," *IEEE Signal Processing Letters,* vol. 23, no. 10, pp. 1499-1503, 2016.

[171] L. Li, X. Feng, Z. Xia, X. Jiang and A. Hadid, "Face Spoofing Detection with Local Binary Pattern Network," *Journal of Visual Communication and Image Representation,* vol. 54, pp. 182-192, 2018.

[172] L.-. B. Zhang, F. Peng, L. Qin and M. Long, "Face Spoofing Detection based on Color Texture Markov Feature and Support Vector Machine Recursive Feature Elimination," *Journal of Visual Communication and Image Representation,* vol. 51, p. 56–69, 2018.

[173] W.-. Y. Zhao, R. Chellappa, P. J. Phillips and A. Rosenfeld, "Face Recognition: A

Literature Survey," *ACM Computing Surveys ,* vol. 35, no. 4, pp. 399-458, 2003.

[174] H.-. T. Nguyen and A. Caplier , "Elliptical Local Binary Patterns for Face Recognition," in *Asian Conference on Computer Vision*, Daejeon, Korea, 2012.

[175] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017.

[176] V. Nair and G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in *27th International Conference on Machine Learning*, Haifa, Israel, 2010.

[177] S. Tirunagari, N. Poh, D. Windridge, A. Iorliam, N. Suki and A. T. S. Ho, "Detection of Face Spoofing Using Visual Dynamics," *IEEE Transactions on Information Forensics and Security,* vol. 10, no. 4, pp. 762 - 777, 2015.

[178] L. Li, Z. Xia, L. Li, X. Jiang, X. Feng and F. Roli, "Face Anti-Spoofing via Hybrid Convolutional Neural Network," in *2017 International Conference on the Frontiers and Advances in Data Science (FADS)*, Xi'an, China, 2017.

[179] Z. Boulkenafet, J. Komulainen and A. Hadid, "On the Generalization of Color Texture-based Face Anti-Spoofing," *Image and Vision Computing,* vol. 77, pp. 1-9, 2018.

[180] F.-. M. Chen1, C. Wen, K. Xie, F. Q. Wen, G. Q. Sheng and X.-. G. Tang, "Face Liveness Detection: Fusing Colour Texture Feature and Deep Feature," *IET Biometrics,* vol. 8, no. 6, pp. 369-377, 2019.

[181] I. Chingovska, A. Anjos and S. Marcel, "On the Effectiveness of Local Binary Patterns in Face Anti-Spoofing," in *International Conference of Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, 2012.

[182] B. Chen, W. Yang and S. Wang, "Face anti-spoofing by fusing High and Low Frequency Features for Advanced Generalization Capability," in *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, Shenzhen, China, 2020.

[183] K. Simonyan and A. Zisserman, "Very Deep Convolutional Newtorks For Large-Scale Image Recognition," in *International Conference on Learning Representations (ICLR)*, San Diego, CA, 2015.

[184] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016.

[185] M. Tan and Q. V. Le , "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *International Conference on Machine Learning*, Long Beach, California, USA, 2019.

[186] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going Deeper with Convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, 2015.

[187] S. Zhang, A. Liu, J. Wan, Y. Liang, G. Guo, S. Escalera, H. J. Escalante and S. Z. Li, "CASIA-SURF: A Large-Scale Multi-Modal Benchmark for Face Anti-Spoofing," *IEEE Transactions on Biometrics, Behavior, and Identity Science,* vol. 2, no. 2, pp. 182 - 193, 2020.

[188] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakanta, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. H.-. Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, "Language Models are Few-Shot Learners," in *34th*

International Conference on Neural Information Processing Systems, Vancouver BC Canada, 2020.

[189] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "An Image is Worth 16x16 Words:Transformers For Image Recognition at Scale," in *International Conference on Learning Representations* , Vienna, Austria, 2021.

[190] G. Wang, C. Lan, H. Han, S. Shan, and X. Chen, "Multi-Modal Face Presentation Attack Detection via Spatial and Channel Attentions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Long Beach, CA, USA, 2019.

[191] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos and S. Marcel, "Biometric Face Presentation Attack Detection With Multi-Channel Convolutional Neural Network," *IEEE Transactions on Information Forensics and Security,* vol. 15, pp. 42 - 55, 2019.

[192] I. Tolstikhin, N. Houlsby, A. Kolesniko, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic and A. Dosovitskiy, "MLP-Mixer: An All-MLP Architecture for Vision," in *35th International Conference on Neural Information Processing Systems*, 2021.

[193] T.-Y. Lin , P. Goyal, R. Girshick , K. He and P. Dollar, "Focal Loss for Dense Object Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 42, no. 2, pp. 318-327, 2020.

[194] A. Antil and C. Dhiman, "A Two Stream Face Anti-Spoofing Framework using Multi-Level Deep Features and ELBP Features," *Multimedia Systems,* vol. 29, p. 1361–1376, 2023.

[195] K. Patel, H. Han and A. K. Jain , "Cross-Database Face Antispoofing with Robust Feature Representation," in *Chinese Conference on Biometric Recognition*, Chengdu, China, 2016.

[196] C. Dhiman, A. Antil, A. Anand and S. Gakhar , "A deep face spoof detection framework using multi-level ELBPs and stacked LSTMs," *Signal, Image and Video Processing,* vol. 18, p. 499–512, 2024.

[197] A. Agarwal, R. Singh and M. Vatsa, "Face Anti-Spoofing using Haralick," in *IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, USA, 2016.

[198] W. Bao, H. Li, N. Li and W. Jiang, "A Liveness Detection Method for Face Recognition based on Optical Flow Field," in *International Conference on Image Analysis and Signal Processing*, Linhai, China, 2009.

[199] Z. Xu, S. Li and W. Deng, "Learning Temporal Features using LSTM-CNN Architecture for Face Anti-spoofing," in *3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, Kuala Lumpur, Malaysia, 2015.

[200] J. Gan, S. Li, Y. Zhai and C. Liu, "3D Convolutional Neural Network Based on Face Anti-spoofing," in *2nd International Conference on Multimedia and Image Processing (ICMIP)*, Wuhan, China, 2017.

[201] D. Li, Y. Yang, Y.-Z. Song and T. M. Hospedales, "Learning to Generalize: Meta-Learning for Domain Generalization," in *AAAI Conference on Artificial Intelligence*, New Orleans Louisiana USA, 2018.

[202] A. Antil and C. Dhiman, "Securing Faces: A GAN-Powered Defense Against Spoofing with MSRCR and CBAM," in *International Conference on Pattern Recognition*, Kolkata, 2024.

[203] J. Lin, C. Gan and S. Han, "TSM: Temporal Shift Module for Efficient Video Understanding," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 2019.

[204] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng and A. Hadid, "OULU-NPU: A Mobile Face Presentation Attack Database with Real-World Variations," in *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2017.

[205] V. J. and S. , "Detecting Pedestrians using Patterns of Motion and Appearance," in *IEEE International Conference on Computer Vision*, Nice, France, 2003.

[206] Y. Atoum, Y. Liu, A. Jourabloo and X. Liu, "Face Anti-Spoofing using Patch and Depth-based CNNs," in *IEEE International Joint Conference on Biometrics (IJCB)*, Denver, CO, USA, 2017.

[207] H.-H. Chang and C.-H. Yeh, "Face Anti-Spoofing Detection based on Multi-Scale Image Quality Assessment," *Image and Vision Computing,* vol. 121, 2022.

[208] Y. Liu, L. Wu, Z. Li and Z. Wang, "Dual-Stream Correlation Exploration for Face Anti-Spoofing," *Pattern Recognition Letters,* vol. 170, pp. 17-23, 2023.

[209] V. L. d. Silva, J. L. Lérida, M. Sarret, M. Valls and F. Giné, "Residual Spatio-Temporal Convolutional Networks for Face Anti-Spoofing," *Journal of Visual Communication and Image Representation,* 2023.

[210] R. Shao, X. Lan, J. Li and P. C. Yuen, "Multi-Adversarial Discriminative Deep Domain Generalization for Face Presentation Attack Detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019.

[211] S. Liu, K.-Y. Zhang, T. Yao, K. Sheng, S. Ding, Y. Tai, J. Li, Y. Xie and L. Ma, "Dual Reweighting Domain Generalization for Face Presentation Attack Detection," in *13th International Joint Conference on Artificial Intelligence (IJCAI)*, Canada, 2021.

[212] Z. Zhang, X. Zhang, C. Peng, D. Cheng and J. Sun, "ExFuse: Enhancing Feature Fusion for Semantic Segmentation," in *European Conference on Computer Vision (ECCV)*, 2018.

[213] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017.

[214] L. v. d. Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research,* vol. 9, pp. 2579-2605, 2008.

[215] "Information Technology-Biometric Presentation Attack Detection—Part 1: Framework," in *document ISO/IEC JTC 1/SC 37 Biometrics*, International Organization for Standardization, 2016.

[216] D. H. Choi, I. H. Jang, M. H. Kim and N. C. Kim, "Color Image Enhancement Based on Single-Scale Retinex With a JND-Based Nonlinear Filter," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, New Orleans, Louisiana, USA, 2007.

[217] D. J. Jobson, Z. Rahman and G. A. Woodell, "A Multiscale Retinex for Bridging the Gap between Color Images and the Human Observation of Scenes," *IEEE Transactions on Image Processing,* vol. 6, no. 7, pp. 965 - 976, 1997.

[218] D. Jobson, Z. Rahman and G. A. Woodell, "Properties and Performance of a Center/Surround Retinex," *IEEE Transactions on Image Processing,* vol. 6, no. 3, pp. 451 - 462, 1997.

[219] S. J. Xie, Y. Lu, S. Yoon, J. Yang and D. S. Park , "Intensity Variation Normalization for Finger Vein Recognition Using Guided Filter Based Singe Scale Retinex," *Sensors,*

vol. 15, no. 7, pp. 17089-17105, 2015.

[220] C.-H. Lee, J.-L. Shih, C.-C. Lien and C.-C. Han, “Adaptive Multiscale Retinex for Image Contrast Enhancement,” in *International Conference on Signal-Image Technology & Internet-Based Systems*, Kyoto,Japan, 2013.

[221] P. Isola, J.-Y. Zhu, T. Zhou and A. A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017.

[222] F. Zhou, C. Gao, F. Chen, C. Li, X. Li, F. Yang and Y. Zhao, “Face Anti-Spoofing Based on Multi-layer Domain Adaptation,” in *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, Shanghai, China, 2019.

[223] R. R. Jingade and R. S. Kunte, “DOG-ADTCP: A new Feature Descriptor for Protection of Face Identification Sytem,” *Expert Systems with Applications,* 2022.

[224] Z. Boulkenafet, J. Komulainen, Z. Akhtar, A. Benlamoudi, D. Samai, S. E. Bekhouche, A. Ouafi, F. Dornaika, A. Taleb-Ahmed, L. Qin, F. Peng, L. B. Zhang, M. Long, S. Bhilare, V. Kanhangad, A. Costa-Pazo, E. Vazquez-Fernandez, D. Perez-Cabo, J. J. Moreira-Perez, D. Gonzalez-Jimenez, A. Mohammadi, S. Bhattacharjee, S. Marcel, S. Volkova, Y. Tang, N. Abe, L. Li, X. Feng, Z. Xia, X. Jiang, S. Liu, R. Shao, P. C. Yuen, W. R. Almeida, F. Andalo, R. Padilha, G. Bertocco, W. Dias, J. Wainer, R. Torres, A. Rocha, M. A. Angeloni, G. Folego, A. Godoy and A. Hadid, “A Competition on Generalized Software-based Face Presentation Attack Detection in Mobile Scenarios,” in *IEEE International Joint Conference on Biometrics (IJCB)*, Denver, CO, USA, 2017.

[225] D. Menotti, G. Chiachia, A. Pinto, W. R. Schwartz, H. Pedrini, A. X. Falcão and A. Rocha, “Deep Representations for Iris, Face, and Fingerprint Spoofing Detection,” *IEEE Transactions on Information Forensics and Security,* vol. 10, no. 4, pp. 864 - 879, 2015.

[226] T. d. F. Pereira, J. Komulainen, A. Anjos, J. M. D. Martino, A. Hadid, M. Pietikäinen and S. Marcel , “Face Liveness Detection using Dynamic Texture,” *EURASIP Journal on Image and Video Processing,* 2014.

[227] H. Li, S. J. Pan, S. Wang and A. C. Kot, “Domain Generalization with Adversarial Feature Learning,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.

[228] H. Liu, F. Liu, X. Fan and D. Huang, “Polarized Self-Attention: Towards High-quality Pixel-wise Regression,” *arXiv:2107.00782 [cs.CV],* 2021.

[229] Y. Feng, F. Wu, X. Shao, Y. Wang and X. Zhou, “Joint 3D face reconstruction and Dense Alignment with Position Map Regression Network,” *arXiv:1803.07835 [cs.CV],* 2018.

[230] S. Fatemifar, M. Awais, S. R. Arashloo and J. Kittler, “Combining Multiple One-Class Classifiers for Anomaly based Face Spoofing Attack Detection,” in *International Conference on Biometrics (ICB)*, Crete, Greece, 2019.

[231] N. Bousnina, L. Zheng, M. Mikram, S. Ghouzali and K. Minaoui, “Unraveling Robustness of Deep Face Anti-Spoofing Models Against Pixel Attacks,” *Multimedia Tools and Applications,* pp. 7229-7246, 2021.

[232] S. Fatemifar, M. Awais, A. Akbari and J. Kittler, “Developing a Generic Framework for Anomaly Detection,” *Pattern Recognition,* 2022.

[233] M. O. Alassafi, M. S. Ibrahim, I. Naseem, R. Alghamdi, R. Alotaibi, F. A. Kateb, H. M. Oqaibi, A. A. Alshdadi and S. A. Yusuf, “A Novel Deep Learning Architecture withImage Diffusion for Robust Face Presentation Attack Detection,” *IEEE Access,*

vol. 11, pp. 59204-59216, 2023.

[234] Y. Wu, D. Tao, Y. Luo, J. Cheng and X. Li, "Covered Style Mining via Generative Adversarial Networks for Face Anti-spoofing," *Pattern Recognition,* vol. 132, 2022.

[235] E. Tzeng,, J. Hoffman, K. Saenko and T. Darrell, "Adversarial Discriminative Domain Adaptation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2019.

[236] Q. Zhou, K.-Y. Zhang, T. Yao, R. Yi, K. Sheng, S. Ding and L. Ma, "Generative Domain Adaptation for Face Anti-Spoofing," in *European Computer Vision Association (ECVA)*, 2022.

[237] Z. Wang, Z. Yu, X. Wang, Y. Qin, J. Li, C. Zhao, X. Liu and Z. Lei, "Consistency Regularization for Deep Face Anti-Spoofing," *IEEE Transactions on Information Forensics and Security ,* vol. 18, pp. 1127 - 1140, 2023.

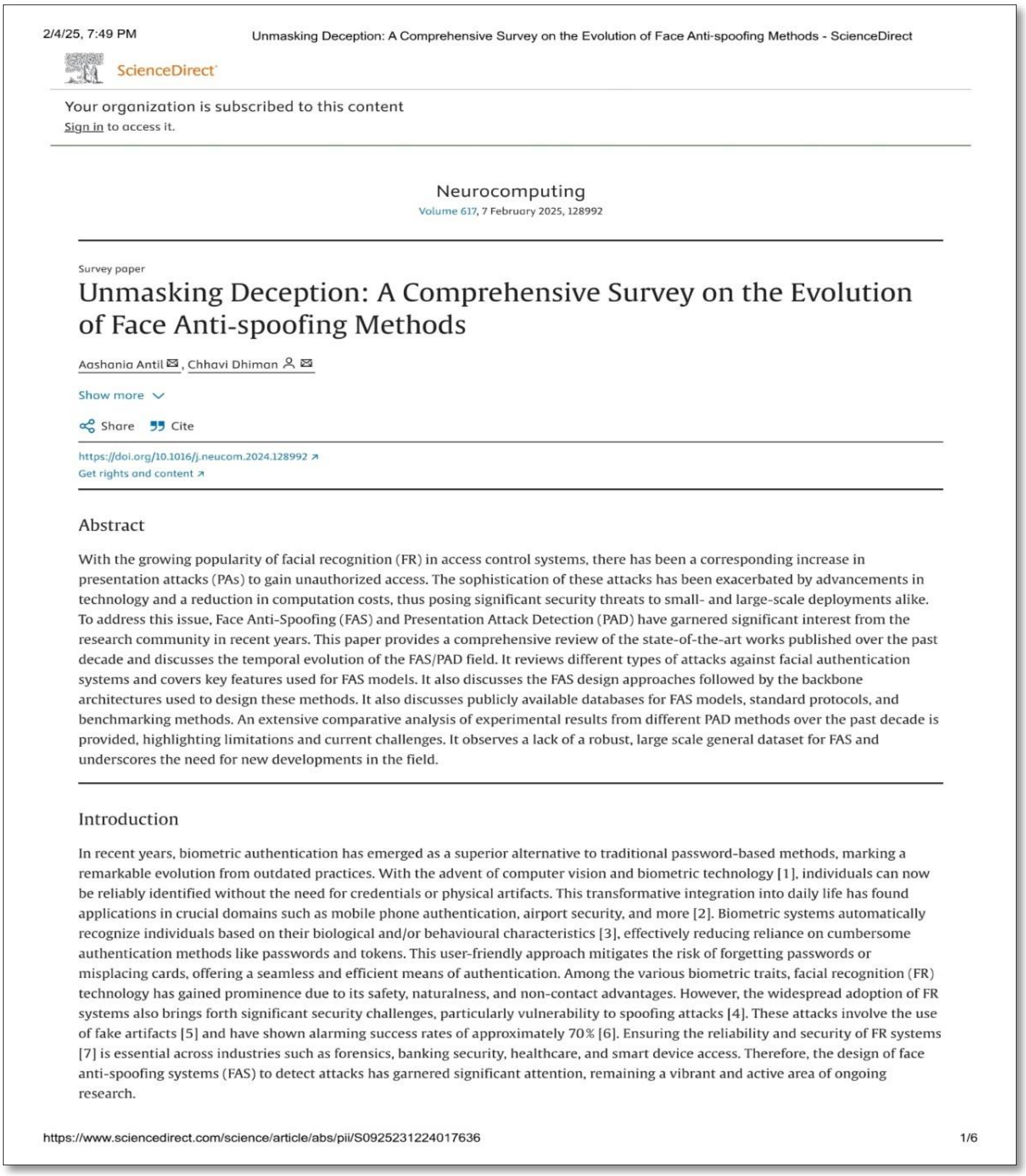
[238] Z. Kong, W. Zhang, T. Wang, K. Zhang, Y. Li, X. Tang and W. Luo, "Dual Teacher Knowledge Distillation with Domain Alignment for Face Anti-spoofing," *arXiv:2401.01102v1,* 2024.

List of Publications Related to the Thesis and their Proofs (document attached)

Papers Published/Accepted for Publication

Journal Papers

- **A. Antil** and C. Dhiman "Unmasking Deception: A Comprehensive Survey on the Evolution of Face Anti-spoofing Methods," ***Neurocomputing***, vol. 617, 2025. SCIE (IF: 5.5). DOI: https://doi.org/10.1016/j.neucom.2024.128992



2/4/25, 7:49 PM Unmasking Deception: A Comprehensive Survey on the Evolution of Face Anti-spoofing Methods - ScienceDirect

ScienceDirect

Your organization is subscribed to this content
Sign in to access it.

Neurocomputing
Volume 617, 7 February 2025, 128992

Survey paper

Unmasking Deception: A Comprehensive Survey on the Evolution of Face Anti-spoofing Methods

Aashania Antil, Chhavi Dhiman

Show more

Share Cite

https://doi.org/10.1016/j.neucom.2024.128992
Get rights and content

Abstract

With the growing popularity of facial recognition (FR) in access control systems, there has been a corresponding increase in presentation attacks (PAs) to gain unauthorized access. The sophistication of these attacks has been exacerbated by advancements in technology and a reduction in computation costs, thus posing significant security threats to small- and large-scale deployments alike. To address this issue, Face Anti-Spoofing (FAS) and Presentation Attack Detection (PAD) have garnered significant interest from the research community in recent years. This paper provides a comprehensive review of the state-of-the-art works published over the past decade and discusses the temporal evolution of the FAS/PAD field. It reviews different types of attacks against facial authentication systems and covers key features used for FAS models. It also discusses the FAS design approaches followed by the backbone architectures used to design these methods. It also discusses publicly available databases for FAS models, standard protocols, and benchmarking methods. An extensive comparative analysis of experimental results from different PAD methods over the past decade is provided, highlighting limitations and current challenges. It observes a lack of a robust, large scale general dataset for FAS and underscores the need for new developments in the field.

Introduction

In recent years, biometric authentication has emerged as a superior alternative to traditional password-based methods, marking a remarkable evolution from outdated practices. With the advent of computer vision and biometric technology [1], individuals can now be reliably identified without the need for credentials or physical artifacts. This transformative integration into daily life has found applications in crucial domains such as mobile phone authentication, airport security, and more [2]. Biometric systems automatically recognize individuals based on their biological and/or behavioural characteristics [3], effectively reducing reliance on cumbersome authentication methods like passwords and tokens. This user-friendly approach mitigates the risk of forgetting passwords or misplacing cards, offering a seamless and efficient means of authentication. Among the various biometric traits, facial recognition (FR) technology has gained prominence due to its safety, naturalness, and non-contact advantages. However, the widespread adoption of FR systems also brings forth significant security challenges, particularly vulnerability to spoofing attacks [4]. These attacks involve the use of fake artifacts [5] and have shown alarming success rates of approximately 70% [6]. Ensuring the reliability and security of FR systems [7] is essential across industries such as forensics, banking security, healthcare, and smart device access. Therefore, the design of face anti-spoofing systems (FAS) to detect attacks has garnered significant attention, remaining a vibrant and active area of ongoing research.

https://www.sciencedirect.com/science/article/abs/pii/S0925231224017636 1/6

- **A. Antil** and C. Dhiman, "A two stream face anti spoofing framework using multi-level deep features and ELBP features," ***Multimedia Systems***, 2023. SCIE (IF: 3.5). DOI: https://doi.org/10.1007/s00530-023-01060-7

Multimedia Systems
https://doi.org/10.1007/s00530-023-01060-7

REGULAR PAPER

A two stream face anti-spoofing framework using multi-level deep features and ELBP features

Aashania Antil[1] · Chhavi Dhiman[1]

Received: 29 November 2022 / Accepted: 5 February 2023

Abstract

The recent boom in publicly available face authentication (FA) system has brought forward the susceptibility of FA systems to different attack vectors, especially spoofing/presentation attacks. There has been a noticeable increase in the novelty, variety and complexity of spoofing attempts on FA systems. Recently, numerous strategies have been proposed, employing both traditional as well as deep learning approaches. However, generalized face spoofing detection has always proven itself as a challenging research hotspot. Here, we propose a multi-level ELBP texture and deep features based novel face anti-spoofing framework for face spoofing detection. Extensive experiments including intra-dataset and inter-dataset testing were performed on three challenging face anti-spoofing databases, namely Replay-Attack, CASIA FASD, and MSU-MFSD to validate the effectiveness of our proposed work. A comprehensive ablation study to analyze the effects of different color spaces and multi-level deep feature extraction is also discussed. The proposed approach is effective at resisting photo and video attacks. The same is confirmed by the experimental results, outperforming the other state-of-arts with competitive results for both intra and inter-dataset evaluation protocols. The code and pre-trained model weights are available at https://github.com/aashania/FAS_Framework_multilevel_ELBP.

Keywords Face authentication · Presentation attacks · Face anti-spoofing · ELBP · Xception

1 Introduction

Face authentication and verification technologies have been deployed in wide array of applications, from access control at borders to face unlocking of smart phones and other daily use devices [1]. High accuracy, fast operation and the general convenience associated with these systems has led to face biometrics' wide spread user adoption. The extensive use of this technology has exhibited vulnerabilities and susceptibility to various forms of attacks such as face spoofing, presentation attacks [2, 3] and face manipulation [4]. The ubiquity of social networks has made it trivial for an attacker to obtain images and videos for face spoofing. This has resulted in bringing the attention of research community toward vulnerability of face authentication-based system. Most proposed countermeasures work adequately in controlled environments. The development of a general anti-spoofing framework which can adapt to device/sensor quality and environment variations has been a challenging task. This is evidenced by the reported drop in performance of state-of-art [5]when exposed to real-world variables such as illumination and background variations, etc. Thus, increasing the effectiveness of face spoofing detection in FR is a crucial task. Therefore, face spoofing detection methods for photo [6] and video attacks [7] have evolved in recent decades.

In the early days, global approaches [8, 9] were the main focus of researchers, in which the entire image is considered as a single vector and the dimensionality of feature vectors was reduced by project into a low dimension sub space with suitable methods e.g., Principal Component Analysis (PCA) [10] and Linear Discriminant Analysis (LDA) [11]. However, when subjected to unrestricted circumstances such as position variations, lighting changes, expression changes, occlusion effects, and aging the works failed to exhibit adequately. Lately, the researchers have been exploring transformation functions such as Patterns of Oriented Edge Magnitudes (POEM) [12], Histogram of Gabor Phase Patterns (HGPP) [13], Local Binary Pattern (LBP) [14], and Local

✉ Chhavi Dhiman
chhavi1990delhi@gmail.com

Aashania Antil
aashiantil40@gmail.com

[1] Delhi Technological University, Delhi, India

Published online: 24 February 2023

- **A. Antil** and C. Dhiman, "MF^{2}ShrT: Multimodal Feature Fusion Using Shared Layered Transformer for Face Anti-spoofing," ***ACM Transactions on Multimedia Computing, Communications, and Applications***, vol. 20, no. 6, pp. 1-21, 2024. SCIE (IF:5.2). DOI: https://doi.org/10.1145/3640817

MF^{2}ShrT: Multimodal Feature Fusion Using Shared Layered Transformer for Face Anti-spoofing

AASHANIA ANTIL and CHHAVI DHIMAN, Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, India

In recent times, Face Anti-spoofing (FAS) has gained significant attention in both academic and industrial domains. Although various convolutional neural network (CNN)-based solutions have emerged, multimodal approaches incorporating RGB, depth, and information retrieval (IR) have exhibited better performance than unimodal classifiers. The increasing veracity of modern presentation attack instruments results in a persistent need to enhance the performance of such models. Recently, self-attention-based vision transformers (ViT) have become a popular choice in this field. Their fundamental aspects for multimodal FAS have not been thoroughly explored yet. Therefore, we propose a novel framework for FAS called MF^{2}ShrT, which is based on a pretrained vision transformer. The proposed framework uses overlap patches and parameter sharing in the ViT network, allowing it to utilize multiple modalities in a computationally efficient manner. Furthermore, to effectively fuse intermediate features from different encoders of each ViT, we explore a T-encoder-based hybrid feature block enabling the system to identify correlations and dependencies across different modalities. MF^{2}ShrT outperforms conventional vision transformers and achieves state-of-the-art performance on benchmarks CASIA-SURF and WMCA, demonstrating the efficiency of transformer-based models for presentation attack detection PAD).

CCS Concepts: • **Computing methodologies;** • **Applied computing** → *Bioinformatics*;

Additional Key Words and Phrases: Face anti-spoofing, presentation attack detection, multimodal, vision transformer

ACM Reference Format:
Aashania Antil and Chhavi Dhiman. 2024. MF^{2}ShrT: Multimodal Feature Fusion Using Shared Layered Transformer for Face Anti-spoofing. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 6, Article 172 (March 2024), 21 pages. https://doi.org/10.1145/3640817

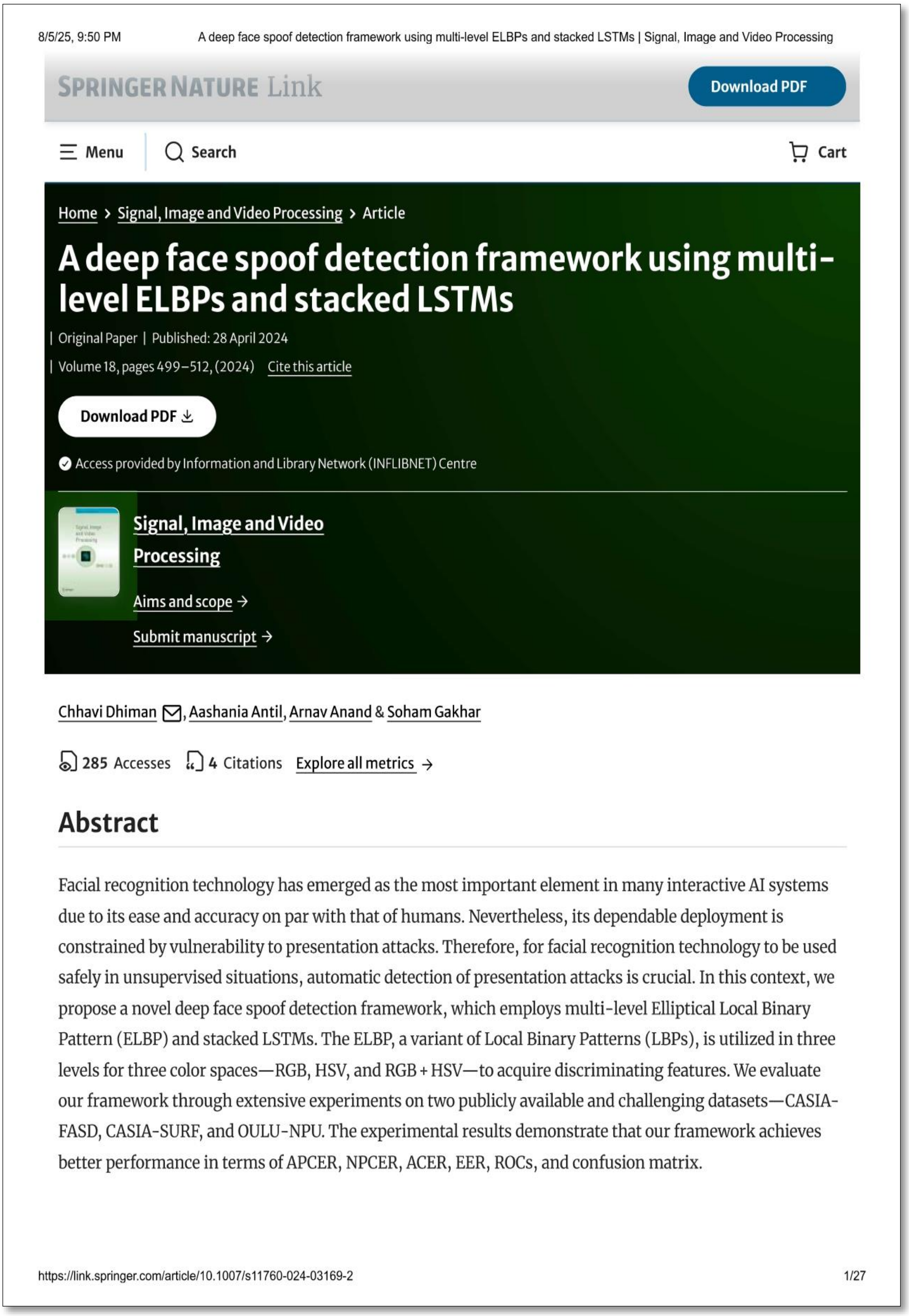
1 INTRODUCTION

Facial recognition technology has become a ubiquitous aspect of modern-day interactive AI systems. Various stakeholders in public and private industries (e.g., airports and banks) employ it in a wide-ranging and ever-growing array of applications, such as electronic payments, security, access control, and surveillance. Despite the advancements in authentication systems, facial recognition technology remains susceptible to **presentation attacks (PAs)** [1], such as print photos [2],

Authors' address: A. Antil and C. Dhiman (Corresponding author), Department of Electronics and Communication Engineering, Delhi Technological University, New Delhi, India; e-mails: aashiantil40@gmail.com, chhavi.dhiman@dtu.ac.in.

ACM 1551-6857/2024/03-ART172
https://doi.org/10.1145/3640817

- C. Dhiman, **A. Antil**, A. Anand and S. Gakhar, “A deep face spoof detection framework using multi-level ELBPs and stacked LSTMs,” *Signal, Image and Video Processing*, vol. 18, p. 499–512, 2024. SCIE (IF:2.0). DOI: https://doi.org/10.1007/s11760-024-03169-2



8/5/25, 9:50 PM
A deep face spoof detection framework using multi-level ELBPs and stacked LSTMs | Signal, Image and Video Processing
SPRINGER NATURE Link
Download PDF
Menu
Search
Cart
Download PDF
Aims and scope →
Submit manuscript →
Explore all metrics →
https://link.springer.com/article/10.1007/s11760-024-03169-2
1/27

Home › Signal, Image and Video Processing › Article

A deep face spoof detection framework using multi-level ELBPs and stacked LSTMs

| Original Paper | Published: 28 April 2024

| Volume 18, pages 499–512, (2024) Cite this article

Access provided by Information and Library Network (INFLIBNET) Centre

Signal, Image and Video Processing

Chhavi Dhiman, Aashania Antil, Arnav Anand & Soham Gakhar

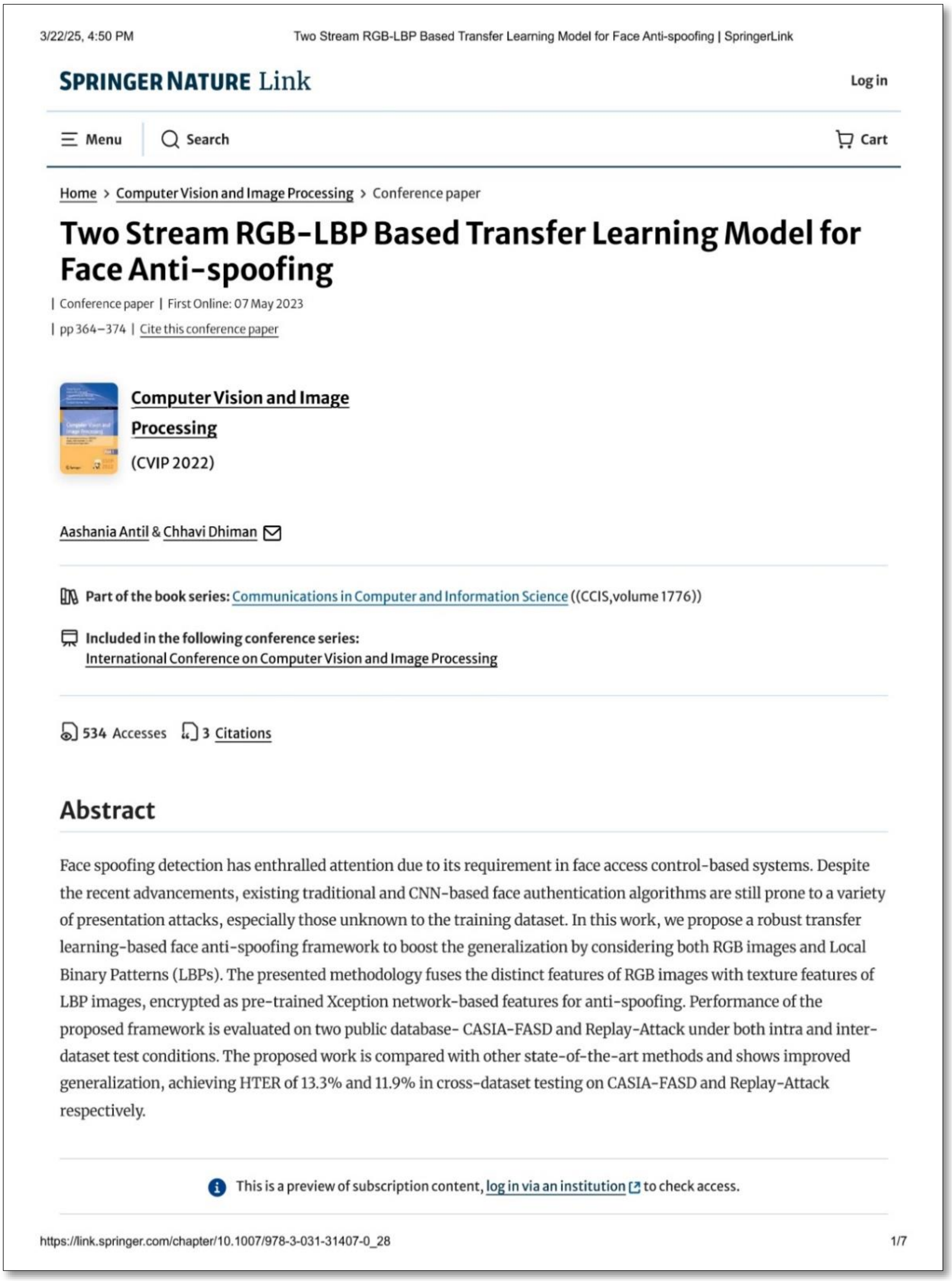
285 Accesses 4 Citations

Abstract

Facial recognition technology has emerged as the most important element in many interactive AI systems due to its ease and accuracy on par with that of humans. Nevertheless, its dependable deployment is constrained by vulnerability to presentation attacks. Therefore, for facial recognition technology to be used safely in unsupervised situations, automatic detection of presentation attacks is crucial. In this context, we propose a novel deep face spoof detection framework, which employs multi-level Elliptical Local Binary Pattern (ELBP) and stacked LSTMs. The ELBP, a variant of Local Binary Patterns (LBPs), is utilized in three levels for three color spaces—RGB, HSV, and RGB + HSV—to acquire discriminating features. We evaluate our framework through extensive experiments on two publicly available and challenging datasets—CASIA-FASD, CASIA-SURF, and OULU-NPU. The experimental results demonstrate that our framework achieves better performance in terms of APCER, NPCER, ACER, EER, ROCs, and confusion matrix.

Conference Papers

- **A. Antil** and C. Dhiman, “Two Stream RGB-LBP Based Transfer Learning Model for Face Anti-spoofing,” in *7th International Conference on Computer Vision & Image Processing (CVIP)*, India, 2023. DOI: https://doi.org/10.1007/978-3-031-31407-0_28


3/22/25, 4:50 PM
Two Stream RGB-LBP Based Transfer Learning Model for Face Anti-spoofing | SpringerLink
SPRINGER NATURE Link
Log in
Menu
Search
Cart
Home > Computer Vision and Image Processing > Conference paper

Two Stream RGB-LBP Based Transfer Learning Model for Face Anti-spoofing

| Conference paper | First Online: 07 May 2023
| pp 364–374 | Cite this conference paper

Computer Vision and Image Processing
(CVIP 2022)

Aashania Antil & Chhavi Dhiman

Part of the book series: Communications in Computer and Information Science ((CCIS,volume 1776))

Included in the following conference series:
International Conference on Computer Vision and Image Processing

534 Accesses 3 Citations

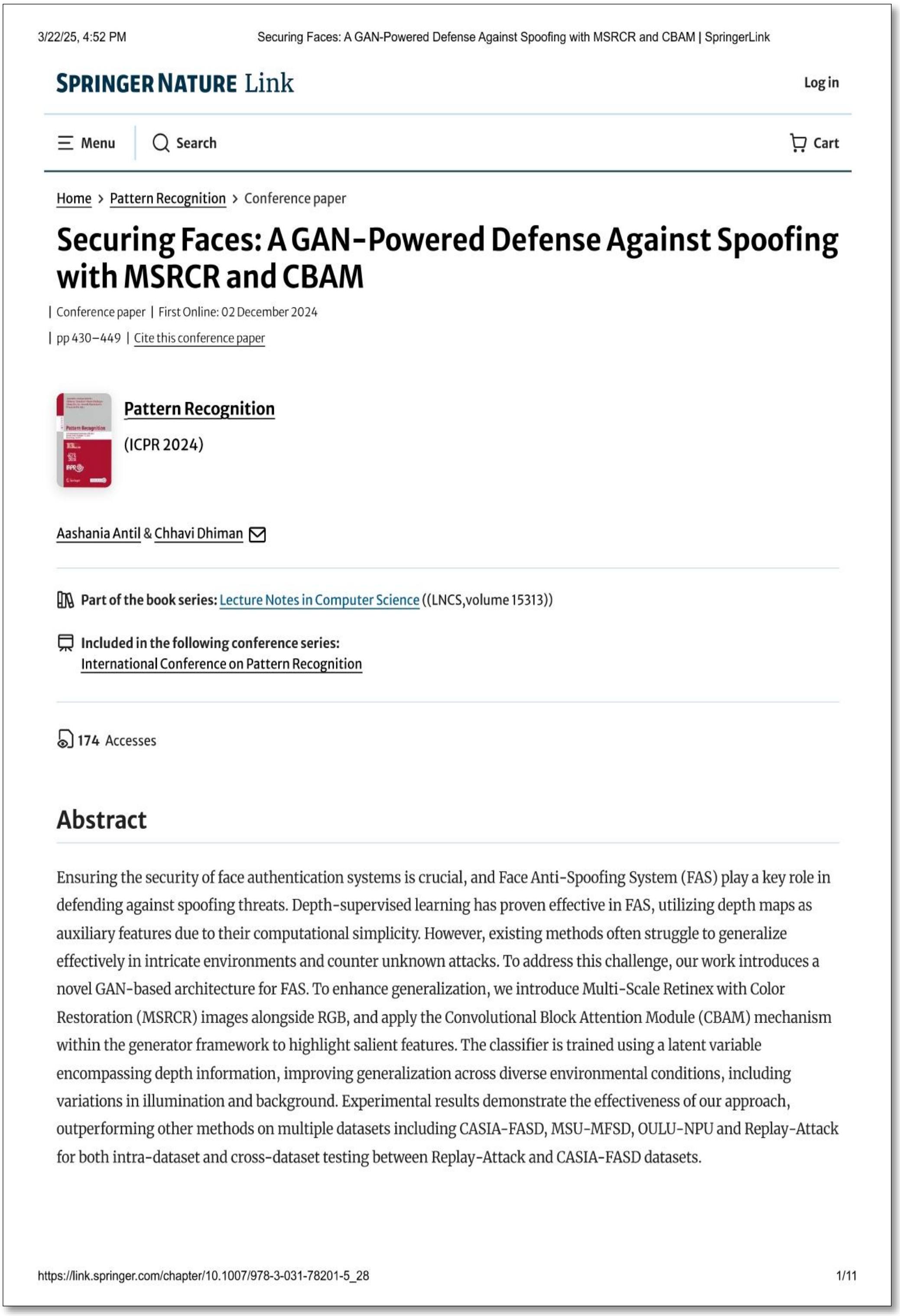
Abstract

Face spoofing detection has enthralled attention due to its requirement in face access control-based systems. Despite the recent advancements, existing traditional and CNN-based face authentication algorithms are still prone to a variety of presentation attacks, especially those unknown to the training dataset. In this work, we propose a robust transfer learning-based face anti-spoofing framework to boost the generalization by considering both RGB images and Local Binary Patterns (LBPs). The presented methodology fuses the distinct features of RGB images with texture features of LBP images, encrypted as pre-trained Xception network-based features for anti-spoofing. Performance of the proposed framework is evaluated on two public database- CASIA-FASD and Replay-Attack under both intra and inter-dataset test conditions. The proposed work is compared with other state-of-the-art methods and shows improved generalization, achieving HTER of 13.3% and 11.9% in cross-dataset testing on CASIA-FASD and Replay-Attack respectively.

This is a preview of subscription content, log in via an institution to check access.

https://link.springer.com/chapter/10.1007/978-3-031-31407-0_28 1/7

- **A. Antil** and C. Dhiman, “Securing Faces: A GAN-Powered Defense Against Spoofing with MSRCR and CBAM,” in 27th International Conference on Pattern Recognition (ICPR), Kolkata, 2024. DOI: https://doi.org/10.1007/978-3-031-78201-5_28



3/22/25, 4:52 PM Securing Faces: A GAN-Powered Defense Against Spoofing with MSRCR and CBAM | SpringerLink

Home > Pattern Recognition > Conference paper

Securing Faces: A GAN-Powered Defense Against Spoofing with MSRCR and CBAM

| Conference paper | First Online: 02 December 2024

| pp 430–449 | Cite this conference paper

Pattern Recognition

(ICPR 2024)

Aashania Antil & Chhavi Dhiman

Part of the book series: Lecture Notes in Computer Science ((LNCS,volume 15313))

Included in the following conference series:
International Conference on Pattern Recognition

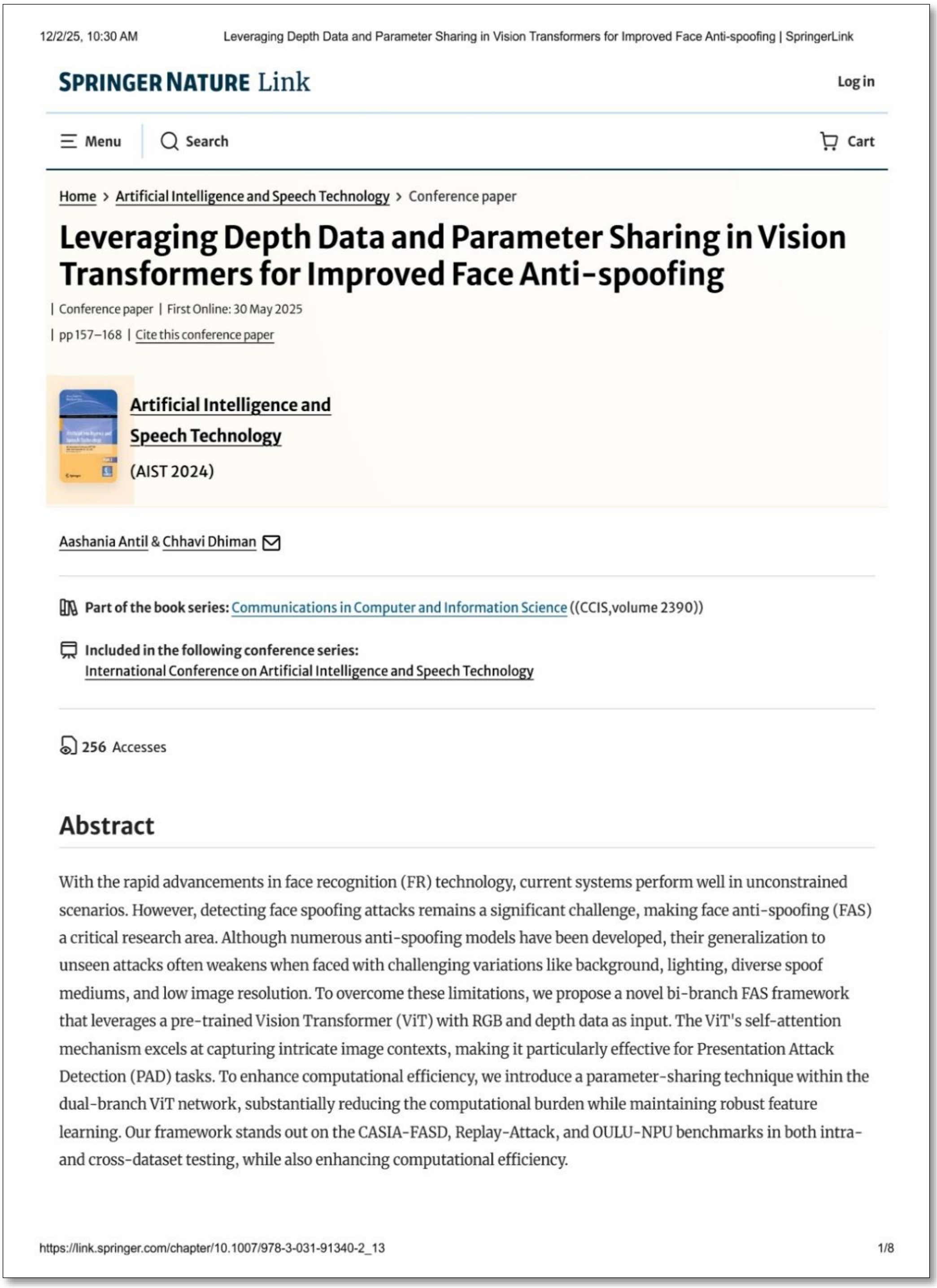
174 Accesses

Abstract

Ensuring the security of face authentication systems is crucial, and Face Anti-Spoofing System (FAS) play a key role in defending against spoofing threats. Depth-supervised learning has proven effective in FAS, utilizing depth maps as auxiliary features due to their computational simplicity. However, existing methods often struggle to generalize effectively in intricate environments and counter unknown attacks. To address this challenge, our work introduces a novel GAN-based architecture for FAS. To enhance generalization, we introduce Multi-Scale Retinex with Color Restoration (MSRCR) images alongside RGB, and apply the Convolutional Block Attention Module (CBAM) mechanism within the generator framework to highlight salient features. The classifier is trained using a latent variable encompassing depth information, improving generalization across diverse environmental conditions, including variations in illumination and background. Experimental results demonstrate the effectiveness of our approach, outperforming other methods on multiple datasets including CASIA-FASD, MSU-MFSD, OULU-NPU and Replay-Attack for both intra-dataset and cross-dataset testing between Replay-Attack and CASIA-FASD datasets.

https://link.springer.com/chapter/10.1007/978-3-031-78201-5_28 1/11

- **A. Antil** and C. Dhiman, "Leveraging Depth Data and Parameter Sharing in Vision Transformers for Improved Face Anti-Spoofing," *6th International Conference on Artificial Intelligence and Speech Technology (AIST)***,** IGDTUW, Delhi, 2024. DOI:https://link.springer.com/chapter/10.1007/978-3-031-91340-2_13



12/2/25, 10:30 AM Leveraging Depth Data and Parameter Sharing in Vision Transformers for Improved Face Anti-spoofing | SpringerLink

Home > Artificial Intelligence and Speech Technology > Conference paper

Leveraging Depth Data and Parameter Sharing in Vision Transformers for Improved Face Anti-spoofing

| Conference paper | First Online: 30 May 2025
| pp 157–168 | Cite this conference paper

Artificial Intelligence and Speech Technology
(AIST 2024)

Aashania Antil & Chhavi Dhiman

Part of the book series: Communications in Computer and Information Science ((CCIS,volume 2390))

Included in the following conference series:
International Conference on Artificial Intelligence and Speech Technology

256 Accesses

Abstract

With the rapid advancements in face recognition (FR) technology, current systems perform well in unconstrained scenarios. However, detecting face spoofing attacks remains a significant challenge, making face anti-spoofing (FAS) a critical research area. Although numerous anti-spoofing models have been developed, their generalization to unseen attacks often weakens when faced with challenging variations like background, lighting, diverse spoof mediums, and low image resolution. To overcome these limitations, we propose a novel bi-branch FAS framework that leverages a pre-trained Vision Transformer (ViT) with RGB and depth data as input. The ViT's self-attention mechanism excels at capturing intricate image contexts, making it particularly effective for Presentation Attack Detection (PAD) tasks. To enhance computational efficiency, we introduce a parameter-sharing technique within the dual-branch ViT network, substantially reducing the computational burden while maintaining robust feature learning. Our framework stands out on the CASIA-FASD, Replay-Attack, and OULU-NPU benchmarks in both intra- and cross-dataset testing, while also enhancing computational efficiency.

https://link.springer.com/chapter/10.1007/978-3-031-91340-2_13 1/8

Communicated

- **A. Antil** and C. Dhiman, “PolarSentinelGAN: A Dual-Polarized Attention-Guided Generative Adversarial Framework for Robust Face Anti-Spoofing,” *IEEE Transactions on Dependable and Secure Computing (TDSC).* SCIE (IF:7.5). (Under Review)
- **A. Antil** and C. Dhiman, “Bi-STAM: Bi-Directional Spatio-Temporal Adaptive Modeling for Robust Face Anti-Spoofing,” *Knowledge-Based Systems*. SCIE (IF:7.6). (Under Review)

DELHI TECHNOLOGICAL UNIVERSITY

Formerly Delhi College of Engineering

Shahbad Daulatpur, Main Bawana Road, Delhi –42

PLAGIARISM VERIFICATION

Title of the Thesis: **Deep Learning Frameworks for Face Ani-spoofing**

Total Pages: **171**

Name of the Scholar: **Aashania Antil**

Supervisor: **Dr. Chhavi Dhiman**

Department: **Electronics and Communication Engineering**

This is to report that the above thesis was scanned for similarity detection. Process and outcome are given below:

Software used: **Turnitin**

Submission ID: **trn:oid:::27535:123598509**

Similarity Index: **30%**

Self-Publication(s) Similarity Index: **21%**

Final Total Similarity Index: **9%**

Total Word Count: **48580**

Date: **December 3, 2025**

Candidate's Signature **Signature of Supervisor**