Enhancing Low-Quality Deepfake Detection Using Hybrid Deep Learning Models

by

Laleh Sadeghi

Bachelor of Computer Science, Azad University, 2011

Master of Public Management, Azad University, 2015

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Computer Science

In the Graduate Academic Unit of Computer Science

Supervisor: Saqib Hakak, Ph.D., Computer Science

Examining Board: Windhya Hansinie Rankothge, Ph.D, Computer Science
Fatemeh Khoda Parast, Ph.D, Computer Science
Clodualdo Aranas, Ph.D, Mechanical Engineering

This thesis is accepted by the Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

February, 2026

© Laleh Sadeghi, 2026

Abstract

The rapid advancement of deep learning–based media synthesis has enabled the creation of highly realistic deepfake videos, posing serious challenges to digital trust and media authenticity. Although existing detection methods often perform well on high-quality data, their effectiveness degrades significantly under real-world conditions where content is repeatedly compressed. This thesis proposes a robust deep learning framework for low-quality deepfake detection using a hybrid feature-fusion architecture. The approach integrates complementary representations extracted by ResNet-50 and EfficientNet-B0, combining fine-grained local texture analysis with global structural and illumination features. To further enhance robustness against compression artifacts, a visibility-based training strategy is incorporated, encouraging the model to focus on subtle forensic traces that persist in degraded media. Experimental evaluations on the FaceForensics++ (C23) and Celeb-DF v2 datasets demonstrate that the proposed hybrid model consistently outperforms individual baseline networks across multiple evaluation metrics, while maintaining computational efficiency.

Dedication

This thesis is dedicated to my beloved daughter, whose love, strength, and patience have been my greatest source of motivation throughout this journey. Her presence gave meaning to every challenge and reminded me why perseverance and hope truly matter.

I also dedicate this work to my family, whose unconditional support, encouragement, and belief in me made this achievement possible. Their sacrifices, understanding, and constant encouragement have been the foundation of my academic and personal growth.

Acknowledgements

I would like to express my sincere and deepest gratitude to my supervisor, Dr. Saqib, for his invaluable guidance, patience, and continuous support throughout the course of this research. His insightful feedback, technical expertise, and constructive suggestions played a crucial role in shaping this thesis and improving the quality of the work at every stage.

I am especially thankful for his encouragement during challenging phases of the research and for providing thoughtful direction that helped me stay focused and motivated. His professionalism and commitment to academic excellence greatly influenced my learning experience and contributed significantly to the successful completion of this thesis.

I would also like to acknowledge the support of the faculty and staff of the University of New Brunswick for providing a stimulating academic environment and the necessary resources for conducting this research.

Finally, I extend my heartfelt appreciation to everyone who supported me throughout this journey, directly or indirectly. Their encouragement and understanding made this achievement possible.

Table of Contents

Abstract ii

Dedication iii

Acknowledgements iv

Table of Contents v

List of Tables ix

List of Figures x

List of Abbreviations xi

Chapter 1 1

Introduction 1

1.1 Introduction 1

1.2 Problem Statement 2

1.3 Research Significance 4

1.4 Research Hypothesis 5

1.5 Research Questions 6

1.6 Research Objectives 6

1.7 Research Contributions 7

1.8 Structure of the Thesis 8

Chapter 2 10

Theoretical Backbone and Literature Review 10

2.1 Historical Development and Evolution of Deepfake Detection 10

2.2 Background 13

2.2.2 Categories of Deepfake Facial Manipulation 17

2.3 Classification of Deepfake Detection Methods 20

2.3.1 Image-Based (Spatial Feature–Based) Methods 20

2.3.2 Temporal and Motion-Based Methods 21

2.3.3 Frequency-Domain Analysis Methods 23

2.3.4 Hybrid and Multimodal Methods 23

2.4 Related Work 24

Chapter 3 27

Methodology 27

3.1 Introduction 27

3.2 Overall Framework of the Proposed Method 27

3.2.1 Input Data and Preprocessing 28

3.2.2 Feature Extraction 31

3.2.3 Visibility Matrix Mechanism 33

3.2.4 Feature Fusion and Reweighting 36

3.2.5 Classification Layer 38

3.3 Summary 39

Chapter 4 41

Simulation and Experimental Results 41

4.1 Introduction 41

4.2 Experimental Setup 41

4.2.1 Datasets 41

4.2.1.1 FaceForensics++ Dataset 42

4.2.1.2 Celeb-DF v2 Dataset 43

4.2.2 Dataset D and DEV Configuration 44

4.2.3 Data Preparation and Frame Extraction 44

4.2.4 Training Configuration 45

4.2.5 Visibility Matrix Training 47

4.2.6 Evaluation Metrics 48

4.3 Baseline Methods and Training Behavior 49

4.3.1 ResNet-50 Baseline 50

4.3.2 EfficientNet-B0 Network 56

4.4 Hybrid Model 58

4.5 Comparison with Baseline Methods 67

4.6 Summary 74

Chapter 5 74

Conclusion and Future Work 74

5.1 Discussion 74

5.1.1 Future Work 75

Bibliography 77

Curriculum Vitae

List of Tables

Table 4-1: Summary of Datasets Used 43

Table 4-2: Performance Comparison of Models..... 69

Table 4-3: Comparison Table: Base Paper vs. Proposed Method..... 71

Table 4-4: Comparison with Recent Deepfake Detection Methods 73

List of Figures

Figure 2.1: Face Swap Method 18

Figure 2.2: Face-to-Face Method 19

Figure 3.1: Overall Framework of the Proposed Method 28

Figure 3.2: Visibility Matrix-Based Learning Mechanism [19] 36

Figure 4.1: Sample training images 46

Figure 4.2: Sample validation images 47

Figure 4.3: Fine-tuning on Dataset D: Training Loss and Validation Accuracy 52

Figure 4.4: Hinge Loss Values During the Training Process of the Visibility Module (per Stage) 53

Figure 4.5: Training Loss Across Different Stages on the Combined Dataset 54

Figure 4.6: Validation Accuracy Across Different Stages on the Combined Dataset 55

Figure 4.7: Fine-tuning on Dataset D using the Hybrid EfficientNet Method 63

Figure 4.8: Hinge Loss Values during the Learning Process of the Visibility Module (per Step) 64

Figure 4.9: Training Loss across Different Stages on the Combined Dataset 65

Figure 4.10: Validation Accuracy across Different Stages on the Combined Dataset 66

List of Abbreviations

Abbreviation	Full Form
AI	Artificial Intelligence
AUC	Area Under the Curve
CNN	Convolutional Neural Network
C23	Compression Level 23 (FaceForensics++ Dataset)
DNN	Deep Neural Network
DF	DeepFake
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
LR	Learning Rate
MBConv	Mobile Inverted Bottleneck Convolution
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SE	Squeeze-and-Excitation
SRM	Spatial Rich Model
VM	Visibility Matrix
FF++	FaceForensics++ Dataset
Celeb-DF v2	Celebrity DeepFake Dataset Version 2
ReLU	Rectified Linear Unit
FC	Fully Connected (Layer)
AdamW	Adam Optimizer with Weight Decay

Chapter 1

Introduction

1.1 Introduction

Deepfake technology, enabled by recent advances in deep learning and generative neural networks, has emerged as a major challenge for digital media authenticity [1,2]. These techniques allow the creation of highly realistic manipulated images and videos that are often indistinguishable from genuine content, even for expert observers [1,2]. As a result, deepfakes pose serious threats to public trust, security, and social stability, particularly when such content spreads rapidly through online platforms [2,3].

A critical challenge in real-world deepfake detection is the degradation of media quality caused by compression and repeated re-encoding on social networks [4-6]. While many existing detection methods achieve strong performance on high-quality or controlled experimental datasets, their effectiveness decreases significantly when applied to low-quality and compressed content [3-8]. Compression processes often suppress or eliminate subtle visual cues that detection models rely on, leading to poor generalization in real-world dissemination scenarios [4,5,9].

Therefore, the central problem addressed in this thesis is the reliable detection of deepfakes under low-quality conditions. Rather than focusing solely on high-resolution data, this research emphasizes realistic settings in which images and videos undergo multiple compression stages [4,6,8]. The goal is to develop deep learning–based detection methods

that remain robust despite quality degradation and are capable of identifying imperceptible artifacts that persist after compression [10-12].

Recent progress in deep neural networks provides promising tools to address this challenge. Modern architectures support multi-scale feature extraction, complementary representations, and enhanced robustness to noise and information loss [13-16]. By leveraging these capabilities, deepfake detection systems can move beyond reliance on fragile, quality-dependent cues and instead learn more stable and generalizable indicators of manipulation [5,17-19].

The significance of this research extends beyond technical considerations. The unchecked spread of manipulated media can damage individual reputations, undermine institutional credibility, and influence public opinion with serious political and social consequences [2,3]. Developing robust and reliable deepfake detection methods is, therefore, not only a scientific objective but also a societal necessity [1,2].

In summary, this thesis focuses on improving the robustness of deepfake detection systems in low-quality and compressed environments. By addressing the gap between controlled experimental conditions and real-world media dissemination, the proposed research aims to contribute toward more practical, reliable, and socially relevant deepfake detection solutions [4-6].

1.2 Problem Statement

With the significant advancements in artificial intelligence and deep learning technologies, the generation of manipulated content such as deepfakes has become one of the

fundamental challenges faced by modern societies [1,2]. Deepfakes, which are typically created using Generative Adversarial Networks (GANs), have enabled the manipulation of images and videos to a level of realism and precision that makes their detection difficult even for domain experts [2,20]. While such content can be employed for positive purposes, including cinematic production and educational applications, its misuse in political, social, and security domains has increased at an alarming rate [7,22].

For instance, fabricated videos of political figures disseminated for the purpose of spreading misinformation can severely undermine public trust and lead to wide-ranging consequences for societal structures [7,17]. One of the most critical challenges in deepfake detection is the impact of data compression. Social media platforms such as Instagram, WhatsApp, and Twitter commonly apply various compression algorithms to reduce data size [4,8]. These compression processes can alter or even remove key features relied upon by deepfake detection algorithms, thereby significantly reducing the accuracy of existing detection systems [4,9,15,23].

Many current deepfake detection approaches are trained on high-quality data; however, these algorithms often fail to maintain acceptable performance when confronted with compressed data [9,15,23]. This limitation highlights the necessity of developing methods capable of operating on invisible or imperceptible features present in compressed content [8,11,12].

One proposed solution is the use of a visibility matrix to facilitate the learning of invisible artifacts in data[19]. This approach, introduced in the base paper underlying this research, aims to identify features that remain detectable even after compression has been applied [19]. By leveraging the visibility matrix, the proposed algorithm guides the model to

enhance and assign greater importance to invisible traces [25,26]. This method not only improves algorithmic performance on high-quality data but also significantly enhances its ability to detect forged content in low-quality data [4,11].

Despite existing efforts, numerous challenges remain in the detection of low-quality deepfakes. These challenges include the selection of effective features, optimization of deep learning algorithms, and evaluation of model robustness under varying data quality conditions [5,27-30]. Furthermore, the requirement for fast and real-time data processing in environments such as social media platforms places additional constraints on the development of these technologies [7].

In addition, limitations such as model sensitivity to natural variations in data, including rotation, illumination changes, and low resolution, pose further obstacles to achieving a fully reliable detection system [1,23]. Therefore, the central problem addressed in this research is the development of algorithms capable of detecting low-quality deepfakes even under compression conditions [4,8,31]. The objective is to design models that are robust to compression and capable of effectively identifying invisible artifacts through techniques such as the visibility matrix [11,25]. This effort not only enhances digital security but also plays a crucial role in combating the spread of misinformation in online environments [2,22].

1.3 Research Significance

This research is significant from scientific, technological, social, and security perspectives, as it addresses the growing challenge of detecting deepfakes in real-world, low-quality

media environments [1,2]. While existing detection methods often rely on high-quality datasets, content shared on social media platforms typically undergoes multiple compression stages that substantially degrade visual cues and reduce detection reliability [4-6,31]. By focusing on robust detection under severe compression, this research addresses a critical gap in current literature and contributes to the development of more practical and generalizable deepfake detection systems [3,7,8,19].

Beyond its scientific value, the work responds to urgent societal needs, as the widespread misuse of deepfakes threatens public trust, individual rights, economic stability, and national security [2,3,22]. The proposed research leverages recent advances in deep neural network architectures to enhance robustness against quality degradation [13-16], thereby supporting more reliable digital forensics tools capable of mitigating legal, ethical, and security risks associated with manipulated media [1,2].

1.4 Research Hypothesis

The present research is based on the following hypotheses:

1. Deep learning–based algorithms can improve the detection accuracy of low-quality deepfakes.
2. The use of frequency-domain features and multi-scale representations increases the robustness of detection algorithms against data compression.
3. The integration of feature reconstruction techniques with quality-aware interactive learning can enhance detection performance across multi-quality datasets.

1.5 Research Questions

This research seeks to address the following questions:

1. How can deep learning–based algorithms improve the accuracy of detecting low-quality deepfakes?
2. How does the use of frequency-domain features and multi-scale representations affect the robustness of detection algorithms against data compression?
3. What impact do feature reconstruction methods and quality-aware interactive learning have on improving detection performance across multi-quality datasets?

1.6 Research Objectives

The primary objectives of this research are as follows:

1. To develop deep learning–based algorithms aimed at improving the detection accuracy of low-quality deepfakes.
2. To evaluate the impact of frequency-domain features and multi-scale representations on enhancing the robustness of detection algorithms against data compression.
3. To design and implement feature reconstruction methods and interactive learning strategies to improve algorithmic performance across multi-quality datasets.

1.7 Research Contributions

The main innovation of this thesis lies in addressing the critical gap between controlled deepfake detection research and real-world media dissemination scenarios, where images and videos are heavily compressed and degraded in quality. While many existing approaches focus primarily on high-quality data, this work explicitly targets low-quality and compressed DeepFake detection, which remains a major challenge in practical deployment.

The key contributions of this thesis are summarized as follows:

- Compression-aware hybrid DeepFake detector:

  A feature-fusion architecture is developed that combines ResNet50 and EfficientNet-B0, enabling the model to jointly exploit fine-grained local textural cues and global structural patterns. This hybrid design improves detection robustness under severe compression while maintaining computational efficiency suitable for practical deployment on low-quality DeepFake data.

- Visibility-matrix–based robustness to compression:

  Motivated by prior work on low-quality deepfake detection using visibility-based learning, this thesis adopts and adapts a visibility-matrix–based training strategy to enhance robustness against compression artifacts. A learnable visibility matrix is applied during training to emphasize subtle forensic traces that may be weakened by compression, guiding feature extraction toward more stable manipulation cues. The proposed framework integrates this strategy into a hybrid ResNet-50 and EfficientNet-B0 feature-fusion architecture and is trained using both original and

visibility-enhanced frames. This adaptation improves robustness to severe compression, illumination variations, and noise compared to single-backbone baseline models.

- Comprehensive multi-quality evaluation protocol:

  A systematic multi-quality evaluation pipeline is constructed using the FaceForensics++ (C23) and Celeb-DF v2 datasets, incorporating realistic frame extraction and compression settings. Performance is assessed using accuracy, balanced accuracy, and ROC-AUC metrics, demonstrating that the developed hybrid model with visibility-based learning consistently outperforms standard CNN baselines while introducing only minimal inference-time overhead.

Overall, this thesis presents a robust, reproducible, and practically applicable framework for low-quality DeepFake detection, contributing both methodological innovations and comprehensive experimental validation under realistic conditions.

1.8 Structure of the Thesis

This thesis is organized into five chapters. Chapter 1 presents the general overview of the research, including the problem statement, the significance and necessity of the study, research objectives, research questions, and research hypotheses.

Chapter 2 reviews the theoretical foundations and related work. This chapter first introduces the fundamental concepts related to deepfakes, generative adversarial networks,

and spatial, temporal, and frequency-domain features, followed by a review of existing approaches in the field of deepfake detection.

Chapter 3 is devoted to the research methodology. This chapter describes the research design and type, the data and experimental datasets, the architecture of the proposed model, the training process, and the evaluation and comparison methods used in this study.

Chapter 4 presents the experimental results obtained from the implementation of the proposed model and compares its performance with that of related models. This chapter provides a detailed analysis of model performance under different quality conditions and real-world scenarios.

Finally, Chapter 5 concludes the thesis by summarizing the main findings, discussing the limitations of the research, and offering recommendations for future work.

Chapter 2

Theoretical Backbone and Literature Review

This chapter presents the theoretical foundation and a comprehensive review of existing research on deepfake detection. It begins by defining key concepts and terminology related to deepfake generation and detection. The chapter then outlines the historical development and evolution of deepfake detection methods, tracing major advances in the field. Finally, existing approaches are systematically categorized and analyzed, with particular emphasis on their strengths, limitations, and effectiveness in low-quality and compressed media scenarios.

2.1 Historical Development and Evolution of Deepfake Detection

The historical development and evolution of deepfake detection can be viewed as a complex and dynamic trajectory that began with the emergence of deep learning–based technologies and has since evolved into one of the most important research areas in computer science and information security [1,2,20]. In the early stages, when the first instances of manipulated content generated by generative networks appeared, the quality of such content was relatively low and exhibited obvious inconsistencies in images and videos [1,21]. For example, lip movements were not synchronized with audio, lighting changes appeared unnatural, and facial regions were poorly integrated with the background [1]. Under these conditions, simple statistical methods and surface-level analyses were sufficient to identify fake content, and deepfakes were not considered a serious threat to security or public trust [13,22].

However, with the rapid advancement of generative algorithms, the quality of deepfakes increased dramatically [7,21]. Within a few years, manipulated content reached a level of realism where distinguishing it from authentic media became extremely difficult [14]. This turning point marked the beginning of extensive efforts to develop more accurate and reliable detection methods [5,10]. In the initial phase of this research domain, investigators primarily relied on approaches based on visual artifact analysis, where algorithms searched for simple inconsistencies in facial components or abnormal eye and lip movements [1]. Although these methods were effective against early generations of deepfakes, they quickly became inadequate as generative technologies improved and resolved many of these visible flaws [10,25,35].

As generative models eliminated such inconsistencies, researchers shifted toward more sophisticated approaches, and deep neural networks emerged as the primary tools for deepfake detection [6,21]. Owing to their ability to learn latent and complex patterns from data, these networks enabled more accurate detection capabilities [14]. During this period, standard benchmark datasets, particularly large-scale deepfake video datasets, were introduced to facilitate model training and evaluation under shared experimental protocols [1,36,37].

A major transformation occurred with the widespread adoption of social media platforms and video-sharing services, where published content became heavily affected by compression algorithms [4,8,37]. This development led to the loss of many key features used for detection, causing existing algorithms to experience substantial performance degradation [4,15,23]. As a result, a new generation of research emerged, focusing specifically on the detection of low-quality deepfakes and compressed data [4,24,25]. At

this stage, researchers recognized the importance of employing multi-scale representations and frequency-domain analysis to uncover subtle forensic cues that persist even after compression [5,10,24].

Consequently, new methods were proposed that incorporated feature reconstruction techniques, joint spatial–temporal analysis, and dual-stream network architectures to improve detection accuracy [5,9,11,15,16]. The evolution of this field demonstrates that each advancement in deepfake generation has been accompanied by a corresponding wave of innovation in detection methods [13]. While simple approaches were sufficient in the early stages, the increasing sophistication of deepfake technology has necessitated the development of more complex and robust algorithms [14,21].

In recent years, the quality of deepfake generation has reached a level that poses serious challenges even for judicial systems and media organizations [7,23]. The historical progression of this domain also highlights the critical role of international collaboration and the creation of large-scale benchmark datasets, which have enabled systematic comparison of methods and accelerated scientific progress [1,36,37]. Furthermore, the historical analysis reveals a growing awareness among researchers that focusing solely on high-quality data is insufficient and that real-world conditions of online content dissemination must be taken into account [4,8]. In practice, users are typically exposed to low-quality, compressed data rather than pristine media [23].

Accordingly, the current phase of deepfake detection research has shifted toward the design of algorithms that are robust under such challenging conditions and capable of operating at scale and in real time [4,11,25,26]. Overall, the history of deepfake detection reflects an ongoing arms race between content generators and detection researchers, where each

improvement in deepfake generation prompts the development of more advanced detection techniques, and vice versa [13]. This continuous interplay has made deepfake detection one of the most dynamic and rapidly evolving research areas in contemporary science [6,21].

With the emergence of recent hybrid approaches that simultaneously leverage multi-scale representations, interactive learning, frequency-domain analysis, and feature reconstruction, there is strong potential for achieving significant improvements in the accuracy and stability of low-quality deepfake detection [16,18,19,24]. This historical trajectory clearly demonstrates that the present research represents a logical continuation of this evolution, aiming to address the limitations of prior studies by proposing a robust and effective algorithm capable of operating reliably under real-world content dissemination conditions and providing a practical response to one of the major challenges of the digital era [17,21].

2.2 Background

In this section, key terms and fundamental concepts related to deepfakes are introduced to provide background knowledge for this research. The section also reviews the main categories of deepfake detection methods, highlighting their core principles and limitations, particularly in low-quality and compressed media scenarios. This background establishes the foundation for the literature review and the proposed methodology presented in later chapters.

2.2.1 Key Terms and Concepts

- Deepfake: Deepfake refers to visual or auditory content that is created or manipulated using artificial intelligence algorithms, particularly deep learning–based models, in such a way that distinguishing between real and fake content becomes extremely difficult for ordinary users and even for experts [1,2,20].
- Low-Quality Video: A low-quality video is defined as a video that has lost a significant portion of its visual details due to compression or resolution reduction, thereby making the detection of manipulated content more challenging [4,8,23].
- Data Compression: Data compression is a process in which the size of images or videos is reduced in order to minimize storage requirements and facilitate data transmission over networks. This process often results in the loss of important information and features, which can negatively affect the accuracy of forgery detection methods [4,10,15,39].
- Deep Neural Network: A deep neural network is a computational model inspired by the structure of the human brain, consisting of multiple layers of artificial neurons. These networks are capable of learning complex data representations and are widely used in image processing and deepfake detection tasks [6,21].
- Multi-Scale Representation: Multi-scale representation refers to the analysis of data at different spatial resolutions and scales, enabling the simultaneous extraction of both fine-grained and coarse features. This technique enhances model robustness against variations in data quality [5,15,16].

- Feature Reconstruction: Feature reconstruction is a method in which information lost due to compression or quality degradation is regenerated, allowing the detection algorithm to make decisions based on more complete and informative data representations [11,27,28].
- Quality-Aware Interactive Learning: Quality-aware interactive learning is a training strategy in which the model is exposed to data with varying quality levels, enabling it to maintain reliable performance under different conditions, including high, medium, and low quality [5,6,25].
- Generative Adversarial Network (GAN): A generative adversarial network consists of two competing networks: one generates synthetic data, while the other discriminates between real and fake data. Through this adversarial process, the quality of generated content improves. GANs are widely used in deepfake creation for generating realistic faces or performing face replacement in videos [2,21].
- Autoencoder: An autoencoder is a model that maps input data to a compact latent representation and then reconstructs the original input from this representation. In classical deepfake generation methods, face reconstruction based on compressed representations enables the transfer of one individual's appearance onto another [1,21].
- Face Alignment: Face alignment is the process of detecting key facial landmarks such as the eyes, nose, and mouth, and transforming the image into a standardized orientation and scale. This step improves training stability and reduces misalignment and distortion during face replacement [1].

- Image Artifacts in Spatial, Temporal, and Frequency Domains: Artifacts are unintended distortions introduced during content generation or editing, such as unnatural edges, abnormal textures, or small visual irregularities. Spatial artifacts appear within individual frames, temporal artifacts emerge across consecutive frames, and frequency-domain artifacts manifest as subtle spectral or texture anomalies [10,14,24,29].
- Compression Artifacts: Compression artifacts include visual distortions such as blockiness, loss of detail, and halo effects around edges that occur as a result of image or video compression. In deepfake content, differences in compression patterns between manipulated regions and authentic areas can serve as important forensic cues [4,8,23].
- Noise and Statistical Residual Models: Noise refers to fine-grained, pseudo-random components introduced by sensors or processing pipelines, which may carry traces of manipulation. Statistical residual models employ specialized filters to emphasize subtle details, allowing manipulated content to be more effectively separated from authentic data [21,24,40].
- Spatial, Temporal, and Frequency Features: Spatial features are extracted from individual frames and include edge shapes, skin textures, and facial proportions. Temporal features analyze changes across frames, while frequency features reveal subtle and abnormal patterns in spectral representations and fine textures [5,9,10,14].

- Diffusion Models: Diffusion models generate images or videos by starting from random noise and progressively denoising it to produce realistic content. Due to their high quality and controllability, these models are increasingly used for generating highly realistic faces and advanced deepfake content [20,22].

2.2.2 Categories of Deepfake Facial Manipulation

There are three main categories used in deepfake detection: face swap, Face2Face, and Reenactment.

I. **Face Swap:** Face swapping is a technique for replacing facial identity, in which the appearance and stable facial characteristics of a source individual are transferred onto the image or video of a target individual, while preserving the target's facial expressions, head movements, lip motion, and blinking behavior [1,2,20]. The process typically begins with data preparation, where the face is detected in each frame, and key landmarks such as the eye corners, nose tip, and mouth contours are extracted for face alignment, ensuring that faces are mapped into a common canonical space [1].

   Subsequently, the facial region is separated from its surroundings, and a precise face mask is generated. A deep learning model is then trained on sufficient samples to learn how to combine the identity of the source with the expressions of the target, producing a face that carries the source identity while retaining the target's facial dynamics [2,20]. After generating the swapped face, a compositing stage is applied, which includes geometric alignment with the target frame, color, illumination, and

shadow matching, as well as seamless edge blending [1]. In video scenarios, temporal consistency techniques are employed to prevent flickering and frame-to-frame instability, resulting in a natural and coherent output [2,41]. Figure 2.1 presents an overview of the face swap process, highlighting the transfer of facial identity from the source to the target while preserving expressions and visual consistency.

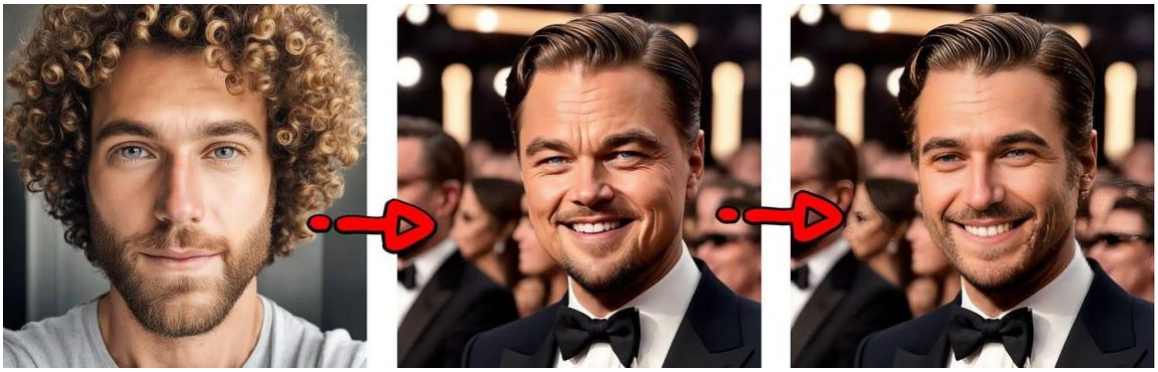

Figure 2.1: Face Swap Method

II. **Face2Face:** Face2Face is a facial reenactment technique originally proposed by Thies et al. [41], which enables real-time transfer of facial expressions and motions from a source face to a target face while preserving the target's identity. Instead of replacing the complete facial identity, this approach transfers controlled facial motions and expressions such as smiling, frowning, lip movement during speech, and blinking from a source face to a target face, while the target retains its original identity [2,41].

The workflow typically starts with face detection in each frame, followed by landmark extraction and head pose estimation to construct a geometric representation of the face [1]. Facial expressions of the source are then modeled as measurable parameters, which are mapped onto the target face so that regions such as the mouth,

cheeks, eyebrows, and eye area change accordingly [7]. To achieve realism, the final composition involves color and illumination correction, boundary smoothing, and temporal stabilization across frames, ensuring smooth transitions and minimizing artificial artifacts [41].

Figure 2.2 illustrates the Face2Face reenactment pipeline, highlighting the expression transfer and compositing stages used to generate the final manipulated output.

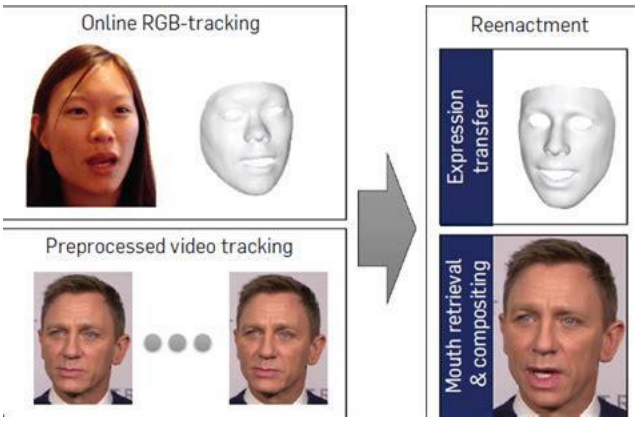

Figure 2.2: Face-to-Face Method

III. **Reenactment**: Reenactment refers to a broader category of deepfake techniques that aim to reproduce facial expressions and head movements of a source individual on a target face, while preserving the target's original identity. Unlike face swapping, reenactment methods do not replace facial identity; instead, they transfer dynamic facial behaviors such as mouth movements, eye blinking, eyebrow motion, and head pose [2,20,41].

Reenactment approaches typically begin with facial tracking in the source video to extract per-frame expression parameters. These parameters are then mapped onto a controllable facial representation of the target, enabling the target face to mimic the expressions and motions of the source in a temporally consistent manner [20,41]. Post-processing steps such as illumination correction, boundary smoothing, and temporal stabilization are commonly applied to reduce visual artifacts and ensure realistic results [2].

2.3 Classification of Deepfake Detection Methods

Deepfake detection methods can be categorized into several major groups, each differing based on the type of data utilized and the analytical approach employed [1,2,20]. This classification highlights the strengths and limitations of individual approaches and clarifies the motivation for hybrid and robust detection frameworks [25,28]. Based on this categorization, deepfake detection methods can be broadly grouped as follows.

2.3.1 Image-Based (Spatial Feature–Based) Methods

Image-based or spatial feature–based methods represent one of the earliest and most extensively studied categories in deepfake detection. These approaches analyze visual information within individual frames to identify artifacts introduced during manipulation, such as inconsistencies in skin texture, illumination, color distribution, and facial geometry [1,21,22,28]. Early studies demonstrated that although deepfake images may appear realistic at a global level, subtle local irregularities often remain detectable through careful

spatial analysis, making image-based features a natural starting point for deepfake detection research [1].

A major line of work within this category focuses on skin texture and illumination analysis, as human skin exhibits fine-grained patterns and lighting behaviors that are difficult for generative models to reproduce accurately [21,28]. In manipulated images, lighting direction, shadows, and reflections may be physically inconsistent across facial regions, while texture patterns may lack natural variability [1]. Other approaches examine the geometric relationships and coordination among facial components, where misalignment between lip movement, eye position, head rotation, and facial expressions can indicate manipulation [1,14,42]. The eye region has also been widely studied, as abnormal blinking behavior, inconsistent pupil motion, or unrealistic corneal reflections frequently appear in deepfake content [1,14].

Recent image-based methods have adopted multi-scale representations, allowing models to analyze both coarse global facial structures and fine-grained local details such as wrinkles and micro-textures [5,12,16]. This strategy improves robustness to resolution loss and compression, which are common in real-world media dissemination scenarios [4,25]. Despite these advancements, image-based methods remain sensitive to severe quality degradation and are often insufficient when used in isolation. Consequently, they are commonly integrated with temporal or frequency-domain approaches to form more robust and generalizable deepfake detection systems [6,8,14,25].

2.3.2 Temporal and Motion-Based Methods

Temporal and motion-based methods analyze dynamic information across consecutive video frames to detect inconsistencies introduced during deepfake generation. Unlike

image-based approaches that focus on single frames, these methods exploit motion continuity, synchronization, and temporal coherence of facial and body movements [5,14]. In authentic videos, facial expressions, eye blinking, lip movements, and head motions evolve smoothly and follow natural physiological patterns, whereas deepfake videos often fail to reproduce such complex temporal behavior accurately [1,9,42].

A major focus of temporal analysis is the examination of eye blinking patterns and lip–audio synchronization. Studies have shown that deepfake videos frequently exhibit abnormal blinking behavior, such as irregular blink frequency or unnatural eyelid dynamics [1,14]. Similarly, mismatches between lip movements and the corresponding speech signal are common indicators of manipulation, as generative models may struggle to achieve precise audiovisual alignment [43,44]. In addition, subtle facial micro-movements that occur naturally during speech or expression changes are often missing or poorly synthesized in manipulated videos, providing further temporal cues for detection [1,9].

Temporal methods also assess frame-to-frame stability and motion correlation among different facial regions and body parts. In real videos, transitions between frames are typically smooth, while deepfake videos may exhibit flickering, abrupt motion changes, or inconsistent coordination between facial motion, head movement, and illumination [4,23]. Although these methods can be affected by low frame rates and heavy compression, recent studies increasingly integrate temporal features with spatial or frequency-domain analysis to improve robustness under real-world dissemination conditions [5,15,24]. As a result, temporal and motion-based methods remain a fundamental component of modern deepfake detection systems, particularly when combined with complementary approaches to enhance generalization and reliability [5,9,15].

2.3.3 Frequency-Domain Analysis Methods

The third category comprises frequency-based approaches, which shift the focus away from visual appearance and instead analyze hidden signals in the frequency domain [10,24]. By examining spectral patterns and fine-grained frequency components, these methods can detect traces left by compression or manipulation processes [8,24]. Frequency-domain analysis is particularly valuable for identifying low-quality deepfakes, as certain forensic cues remain detectable even after aggressive compression [4,8,25,29].

2.3.4 Hybrid and Multimodal Methods

The fourth category includes hybrid and multimodal approaches, which integrate multiple data sources or representations to enhance detection performance [5,15,16]. These methods may combine audio and visual data or fuse multi-scale spatial, temporal, and frequency representations within a unified framework [17,24,44]. By leveraging complementary information from different domains, hybrid approaches often achieve higher robustness and generalization compared to single-modality methods [5,11,14,25].

Overall, this classification demonstrates that while each approach possesses inherent strengths and weaknesses, their integration can provide a more comprehensive and effective solution [13,28]. In recent years, research has increasingly focused on multi-layer and compression-robust methods capable of accurately detecting low-quality deepfakes [48,25,26]. This direction continues to represent a central and active line of investigation in the field [11,13,22].

2.4 Related Work

Recent advances in deepfake detection have increasingly shifted toward addressing robustness under low-quality and compressed conditions, motivated by the reality that online media is routinely degraded through social media dissemination pipelines. Early deepfake detection methods primarily relied on spatial-domain analysis of individual frames, exploiting inconsistencies in facial texture, illumination, and geometry [1,13,21]. As generative models improved, researchers proposed more sophisticated spatial architectures, including multi-scale and dual-stream networks, to capture both local and global manipulation cues. Cheng et al. introduced a multi-scale interactive dual-stream network to enhance detection performance on degraded videos [13], while Han et al. incorporated learnable SRM filters within a two-stream framework to expose subtle artifacts that persist after compression [21]. Attention-based models further improved robustness by dynamically focusing on informative facial regions, as demonstrated by Zhao et al. through multi-attentional learning [14] and by Lee et al. with BZNet, an unsupervised multi-scale branch zooming network designed for low-quality deepfake videos [15]. Vision transformer–based and hybrid CNN–ViT architectures have also been explored to improve representation robustness under visual degradation [16].

Beyond spatial analysis, a growing body of work has emphasized frequency-domain representations, feature restoration, and cross-quality learning as effective strategies for mitigating compression effects. Gao et al. proposed a high-frequency enhancement network that amplifies suppressed spectral components to improve detection on highly compressed content [10], while Wang et al. leveraged frequency-domain filtered residual networks to capture manipulation traces in low-quality data [8]. Additional studies have

shown that frequency and phase-based features remain informative even when spatial cues are heavily degraded [29]. Feature restoration techniques have also been investigated, including a spatial restoration framework for social media deepfake videos proposed by Li et al. [4] and DF-UDetector, introduced by Ke and Wang, which explicitly reconstructs compression-degraded features [11]. Knowledge distillation has emerged as another powerful tool for cross-quality generalization, with Kim et al. proposing FReTAL to improve robustness under distribution shifts [17], and Woo introducing frequency-attention and multi-view knowledge distillation specifically targeting low-quality compressed deepfake images [18]. Recent studies further emphasize quality-agnostic and cross-quality learning strategies to improve performance stability across varying compression levels [5,16,19].

Despite these advances, several limitations remain unresolved in existing deepfake detection research. Many methods still exhibit sensitivity to severe quality degradation, overfitting to dataset-specific artifacts, and reduced generalization across different codecs, resolutions, and real-world compression pipelines [3,5-7]. Furthermore, a large portion of the literature focuses on isolated feature domains, spatial, temporal, or frequency, without fully exploiting their complementary nature within a unified framework [16,25]. Cross-quality evaluation and robustness under repeated compression stages are also often insufficiently addressed [4,8,23]. Recent work on compression-invariant representations highlights the importance of learning stable forensic cues that persist across quality variations [26]. These limitations motivate the development of hybrid detection frameworks that jointly integrate robust spatial representations, compression-aware learning, and invariant feature modeling. The proposed approach in this thesis directly

addresses these challenges by combining hybrid feature fusion with visibility-matrix–based learning to achieve reliable deepfake detection under realistic low-quality and compressed conditions.

Chapter 3

Methodology

3.1 Introduction

In this chapter, the methodology of the proposed low-quality deepfake detection approach is presented. The chapter outlines the overall framework of the proposed system, encompassing data preprocessing, feature extraction, the visibility matrix–based learning strategy, and model training. The datasets used in this study, along with the evaluation metrics employed to assess detection performance under various quality conditions, are also introduced. This chapter provides the technical foundation required for the experimental analysis presented in Chapter 4.

3.2 Overall Framework of the Proposed Method

The overall framework of the proposed low-quality deepfake detection method is designed as a structured, end-to-end processing pipeline that explicitly addresses the challenges posed by compression, resolution degradation, and real-world content distribution conditions [4-6]. An overview of the complete framework is illustrated in Figure 3.1, while a detailed representation of the visibility matrix integration is shown in Figure 3.2.

The proposed framework integrates multiple complementary stages, including input preparation, multi-scale feature extraction, visibility-based feature evaluation, quality-aware feature reweighting, and final classification. Together, these components form a coherent and robust detection pipeline tailored for low-quality deepfake scenarios.

Feature extraction is performed using complementary backbone networks (ResNet-50 and EfficientNet-B0) to capture both fine-grained local texture patterns and global structural representations. To enhance robustness against compression artifacts, a visibility matrix–based learning mechanism, originally proposed in [19], is incorporated into the framework. In this work, the mechanism is integrated within the hybrid architecture to evaluate feature stability under quality degradation and to adaptively suppress unstable features while enhancing robust manipulation cues prior to classification.

The final classifier operates on the reweighted feature representation to produce binary predictions (real vs. deepfake).

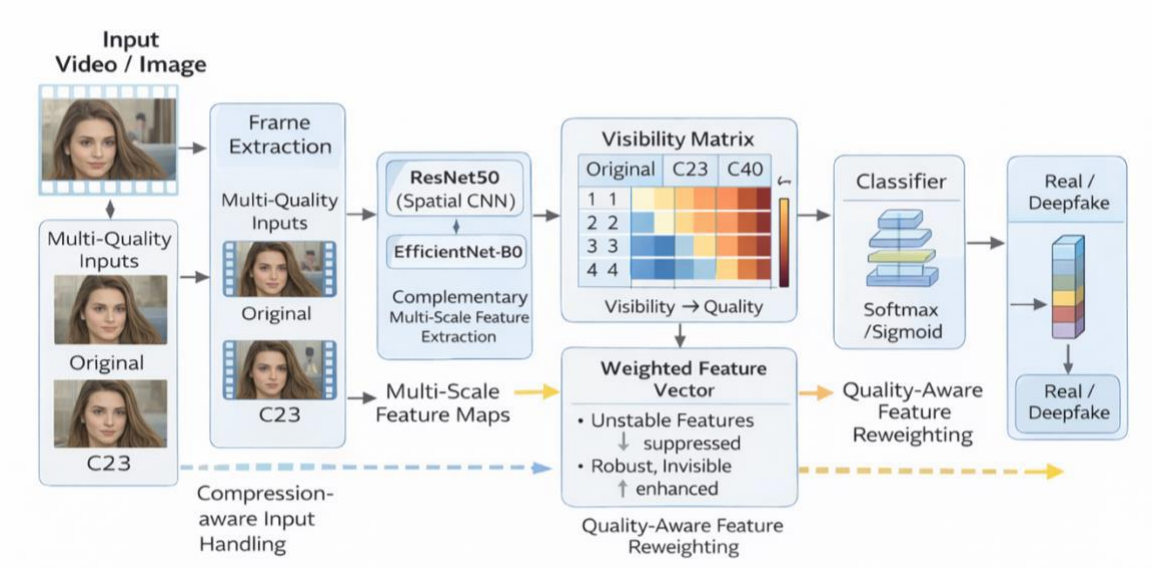

Figure 3.1: Overall Framework of the Proposed Method

3.2.1 Input Data and Preprocessing

Input data preparation is a critical component of the proposed low-quality deepfake detection framework, as the stability, robustness, and generalization capability of the model

are directly influenced by the quality and consistency of the input data [1,2]. In real-world scenarios, deepfake content is typically distributed through online platforms where compression, resolution degradation, and transmission artifacts are unavoidable [4-6]. Consequently, preprocessing in this research is designed not merely as a standard normalization step, but as an integral part of the methodology aimed at reflecting realistic content distribution conditions while maintaining controlled and reproducible input representations.

The input data to the proposed system consists of videos and images obtained from publicly available deepfake datasets [4,37]. Since direct video processing is computationally expensive and may introduce redundancy, video data are first converted into a set of representative image frames [13-15]. Frame extraction follows a fixed and reproducible sampling policy, where a predefined number of frames is selected from each video at controlled temporal intervals [6,25]. This strategy preserves diversity in facial expressions, illumination conditions, and motion patterns while keeping the dataset size computationally manageable.

Following frame extraction, an initial filtering stage is applied to manage severely degraded samples. Frames exhibiting extreme blur, heavy occlusion, missing facial regions, or insufficient visual information are either excluded from the dataset or incorporated with reduced weighting, depending on the predefined policy [4,7]. This step prevents severely corrupted samples from destabilizing the learning process while still allowing the system to encounter realistic degradation scenarios.

Next, face detection and region-of-interest (ROI) extraction are performed to constrain the model's focus to the relevant facial area [42,44]. Facial regions are detected and cropped

to remove background interference and reduce irrelevant variability. To further stabilize spatial alignment, facial landmark detection is employed to align faces before cropping [42]. This alignment ensures consistent positioning of key facial components, such as the eyes, nose, and mouth, reducing unnecessary geometric variation and allowing the model to allocate its learning capacity to manipulation-related cues rather than pose or rotation differences.

After ROI extraction, spatial resizing is applied to standardize all input samples to a fixed resolution [33,34]. This step ensures uniform input dimensionality across the dataset and enables meaningful feature-space comparisons. The selected resolution balances the preservation of essential visual details with computational efficiency, avoiding unnecessary increases in training cost.

Intensity and color normalization are subsequently performed to reduce variations caused by camera properties, lighting conditions, and recording environments [34]. Pixel values are mapped to a standardized range, and normalization procedures are applied where appropriate. These operations encourage the model to focus on structural and statistical inconsistencies associated with manipulation rather than superficial color or illumination differences.

To align preprocessing with the primary objective of low-quality deepfake detection, controlled quality degradation scenarios are explicitly introduced [4,6,8]. Multiple versions of the input data are generated under different levels of compression and resolution reduction to simulate realistic digital content distribution pipelines. This multi-quality data preparation strategy exposes the model to a spectrum of quality conditions during training and evaluation, discouraging reliance on fragile visual cues that disappear under

compression and promoting the learning of stable and persistent manipulation artifacts [10,11].

In addition to quality degradation, limited and non-destructive data augmentation techniques are applied when necessary to improve generalization [27,30]. These augmentations include minor variations in brightness and contrast, small spatial translations, and low-degree rotations. All augmentation operations are carefully constrained to preserve manipulation-related characteristics and avoid introducing unrealistic distortions that could interfere with feature learning.

Class balance management is also incorporated into the preprocessing pipeline. Deepfake datasets often exhibit an imbalance between real and fake samples or unequal diversity across classes [5,7].To mitigate bias toward dominant classes and reduce critical misclassification errors, balanced sampling strategies or class-weighted loss functions are employed during training [38].

3.2.2 Feature Extraction

Feature extraction constitutes the technical core of the proposed low-quality deepfake detection system, as the discriminative capability and robustness of the model are directly determined by the quality of the learned feature representations [1,2]. If the extracted features are insufficiently informative or overly sensitive to quality degradation, the detection performance deteriorates significantly, particularly in real-world scenarios involving compression and resolution loss [4-6]. Consequently, the feature extraction

strategy in this research is designed to capture stable and complementary cues that remain effective under varying quality conditions.

In this work, feature extraction is performed using deep convolutional neural networks rather than manually designed features. Specifically, two backbone architectures, ResNet-50 and EfficientNet-B0, are employed to extract spatial feature representations from input frames. These networks are selected due to their complementary architectural characteristics and their proven effectiveness in learning discriminative visual representations under challenging conditions [10,13,14].

The ResNet-50 branch focuses on capturing fine-grained local texture inconsistencies and residual artifacts introduced by deepfake manipulation, such as blending irregularities and boundary distortions. Features are extracted from the final convolutional layer of ResNet-50 and aggregated using global average pooling, resulting in a 2048-dimensional feature vector. In parallel, the EfficientNet-B0 branch captures broader structural and contextual facial patterns using compound scaling and squeeze-and-excitation mechanisms. After adaptive average pooling, this branch produces a 1280-dimensional feature vector that emphasizes global facial geometry and structural consistency.

At a conceptual level, each dimension of the extracted feature vector represents the activation strength of a learned visual pattern within the image. These learned patterns may correspond to texture discontinuities, edge irregularities, unnatural blending boundaries, illumination asymmetries, or compression artifacts introduced during deepfake synthesis. In practical terms, each feature value indicates how strongly a specific manipulation-related cue is present in the input frame. Therefore, the feature vector acts as a compact

numerical summary of meaningful visual evidence that supports the final classification decision.

Although no explicit handcrafted frequency-domain features are computed, frequency-related artifacts are implicitly encoded within the learned deep representations. Convolutional filters in the backbone networks are sensitive to high-frequency residuals and compression artifacts that persist even after aggressive re-encoding [8,24,25]. Additionally, temporal and motion-related inconsistencies are learned implicitly through frame-level training on video data under multiple compression settings, enabling the model to capture frame-to-frame instability without relying on explicit optical flow computation [6,10,11].

The extracted feature vectors from both backbone networks are subsequently concatenated and passed to the visibility-matrix-based feature reweighting module described in Section 3.2.3. This fusion allows the system to jointly exploit local texture cues, global structural patterns, and compression-resilient residual information, providing a robust feature representation suitable for low-quality deepfake detection.

3.2.3 Visibility Matrix Mechanism

Figure 3.2 illustrates the visibility matrix–based learning mechanism adopted in this work. The visibility matrix concept was originally proposed in [19] to enhance robustness against compression artifacts in low-quality deepfake detection.

In the proposed framework, this mechanism is integrated to evaluate feature stability under quality degradation and to adaptively suppress unstable features while enhancing robust and imperceptible manipulation cues prior to final classification. By incorporating this

strategy within the hybrid architecture, the model improves generalization under low-quality and compressed conditions.

The visibility matrix module represents a central component of the proposed low-quality deepfake detection framework. Its primary role is to evaluate and regulate the contribution of extracted features under varying quality conditions, particularly in the presence of compression and resolution degradation [4-6]. In low-quality scenarios, many discriminative visual cues become weakened or imperceptible, while others remain relatively stable [8,10]. The visibility matrix is designed to explicitly model this behavior and guide the learning process toward robust and quality-invariant features [1,2].

Within the proposed framework, the visibility matrix operates after feature extraction and before final classification. It functions as a learnable mechanism that analyzes feature responses across different quality levels and assesses their relative stability [5,8]. Features that remain consistent and informative despite quality degradation are assigned higher importance, whereas features that are highly sensitive to compression artifacts or resolution loss are suppressed [4,6]. This selective emphasis enables the system to reduce reliance on fragile cues that are likely to disappear in real-world content distribution scenarios [3,5].

The motivation behind introducing the visibility matrix stems from the observation that, although compression degrades visible details, residual traces of manipulation often persist in the data [10,11]. These traces may not be visually apparent but can still be detected by learning-based systems if appropriately amplified [17,18]. The visibility matrix facilitates this process by enhancing subtle and latent features that are otherwise difficult to exploit, particularly in heavily compressed inputs [8,10].

The visibility matrix is incorporated as a learnable component within the training pipeline. During training, the module evaluates feature behavior under multiple quality conditions and progressively adapts its weighting strategy [4,6]. This enables the model to focus on deeper and more stable patterns rather than superficial, quality-sensitive features [1,5]. As a result, the learning process becomes explicitly quality-aware, encouraging robustness across different compression levels [3,8].

In the proposed hybrid architecture, the visibility matrix interacts closely with the feature reweighting mechanism. Its output serves as guidance for determining how individual feature channels or groups should be emphasized or suppressed [6,13]. This interaction ensures that feature reweighting is not performed arbitrarily but is instead informed by measured feature stability under degradation [4,5].

The functional role of the visibility matrix within the proposed method can be summarized as follows: it measures feature stability under compression, suppresses unstable and fragile features, enhances invisible but robust features, and supports quality-aware learning [6,8,10]. By integrating these functions into the feature processing pipeline, the visibility matrix contributes directly to improving cross-quality generalization and detection reliability [3,5].

Overall, the visibility matrix module enables the proposed system to move beyond conventional feature extraction by explicitly accounting for the impact of quality degradation on feature visibility [4,6]. Through its learnable and adaptive design, the module plays a key role in guiding the model toward robust representations that remain effective under low-quality and real-world conditions [1,2].

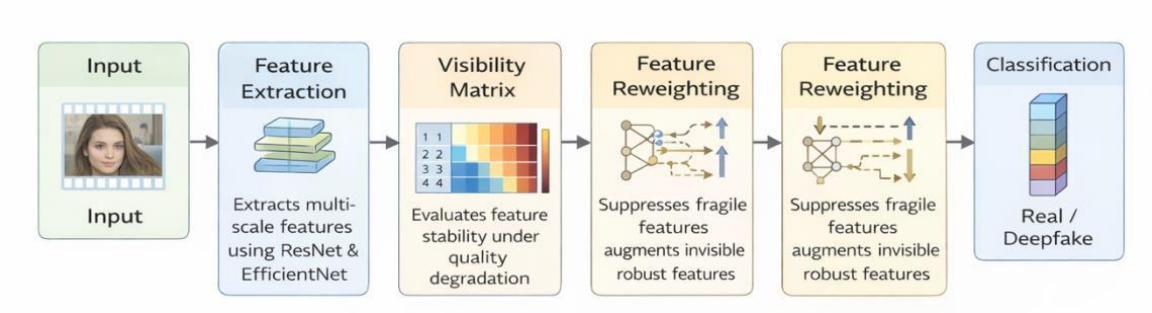

Figure 3.2: Visibility Matrix-Based Learning Mechanism [19]

3.2.4 Feature Fusion and Reweighting

After spatial, temporal, and frequency-domain features are extracted, an effective fusion and reweighting strategy becomes essential to form a coherent and robust representation [1,2]. Each feature domain provides a distinct perspective on deepfake manipulation, and an unstructured or naive combination of these features may lead to interference, dominance of one modality, or dilution of useful information [16,25]. Therefore, feature fusion in this research is carefully designed to preserve complementarity while adapting to varying quality conditions [13,14].

The fusion process goes beyond the simple concatenation of feature vectors. Instead, the extracted features are first aligned to ensure comparability across different representation spaces [8,24]. This alignment normalizes scale and statistical properties, reducing conflicts between features learned from different domains and network branches [1]. By enforcing consistency at this stage, the system prevents unstable feature interactions that could negatively affect learning and generalization [5,6].

Following alignment, feature reweighting is applied to regulate the relative contribution of each feature group. The reweighting mechanism is designed to adaptively emphasize features that remain stable and informative under quality degradation while suppressing those that are fragile and highly sensitive to compression artifacts [4,10,11]. This behavior is particularly important in low-quality scenarios, where certain spatial cues may be weakened or partially lost, while temporal or frequency-based cues may retain higher discriminative value [8,25].

The reweighting process is quality-aware and dynamically adjusts feature importance based on the reliability of the extracted cues [3,5]. Under higher-quality conditions, fine-grained spatial features may contribute more significantly to the final representation. As quality degrades, the system can shift its reliance toward more robust temporal or frequency-domain features [8,10]. This adaptive behavior reduces dependence on superficial cues and encourages the model to focus on stable manipulation traces [6].

In the proposed hybrid architecture, feature fusion and reweighting are tightly coupled with the visibility matrix module. The visibility matrix guides feature stability under different quality conditions, and this information directly influences how features are weighted during fusion [4,5]. As a result, the final representation reflects not only the presence of discriminative features but also their reliability across compression levels [3,6].

Through this structured fusion and reweighting strategy, the proposed system constructs a shared representation that integrates complementary information from multiple domains while maintaining robustness to quality degradation [1,13]. By dynamically regulating feature contributions, the model reduces sensitivity to resolution loss, compression artifacts, lighting variations, and pose changes [4-6]. This design plays a crucial role in

enabling consistent deepfake detection performance across diverse and challenging real-world conditions [2,3].

3.2.5 Classification Layer

Following feature extraction, visibility-based evaluation, and adaptive fusion, the final feature representation is passed to the classification head, which is responsible for producing the ultimate decision regarding the authenticity of the input content [1,2] This stage maps the learned and fused features into the binary decision space of real versus fake.

Before classification, the outputs of the feature extraction branches are processed through global average pooling layers [32,33]. This operation converts high-dimensional spatial feature maps into compact feature vectors. Global pooling offers two key advantages: it significantly reduces the number of parameters in the classification stage, thereby lowering the risk of overfitting, and it decreases sensitivity to precise spatial locations of features [32,34]. As a result, the learned representations become more robust to resolution changes and quality variations [5,6].

The pooled feature vectors are then combined to form a shared representation that integrates information from multiple processing pathways [13,14]. This combined feature vector is subsequently fed into one or more fully connected layers that act as the decision head [33]. These layers are designed to transform the fused representation into a discriminative space suitable for classification [1].

To promote generalization and prevent overfitting, regularization strategies are applied within the classification head [34]. These include controlled layer dimensionality and dropout mechanisms, which encourage the model to learn stable and transferable

discrimination patterns rather than memorizing training data [32,33]. Such regularization is particularly important in low-quality deepfake detection, where noise and compression artifacts can otherwise lead to unstable decision boundaries [4-6].

The output of the classification head is generated either as a probability score or as a final class label, representing the likelihood that the input content is real or fake [38]. This output serves as the basis for performance evaluation under different quality conditions and forms the final stage of the proposed detection framework [2,3].

3.3 Summary

In summary, the research methodology of this thesis was designed in a structured, coherent, and step-by-step manner to address the problem of low-quality deepfake detection under conditions close to real-world content distribution [4-6]. Initially, datasets were selected from well-established and widely used sources [1,2], and the data were prepared through representative frame extraction, removal of unsuitable samples, face detection and cropping, spatial normalization, and intensity normalization [4,37].

To realistically reflect content dissemination scenarios on social media platforms, controlled quality degradation and compression scenarios were applied [3,4,8], enabling the proposed system to learn robust representations and reduce sensitivity to resolution loss [5,6]. Subsequently, a comprehensive feature extraction strategy was developed, focusing on spatial, temporal, and frequency-domain cues, followed by their purposeful integration to capture complementary information relevant to deepfake detection [1,13,14].

The proposed model was trained using a hybrid architecture combined with appropriate regularization mechanisms to mitigate overfitting and enhance generalization capability [32-34]. The training process was carefully managed through continuous monitoring of validation data and the application of early stopping strategies to ensure stable and reliable convergence [34].

Finally, system performance was evaluated using a diverse set of evaluation metrics across multiple quality levels [38], allowing a reliable assessment of robustness and effectiveness under challenging low-quality conditions. Overall, this chapter establishes a solid methodological foundation that supports the experimental analysis and results presented in the subsequent chapter.

Chapter 4

Simulation and Experimental Results

4.1 Introduction

This chapter presents the experimental evaluation of the proposed deepfake detection framework. ResNet-50 and EfficientNet-B0 are used as baseline feature extractors, and a hybrid model is developed to improve detection accuracy and robustness under realistic low-quality conditions. Experiments are conducted on FaceForensics++ (C23) and Celeb-DF v2 using extracted video frames resized to 128 × 128 pixels. The results are reported using accuracy and additional evaluation measures, and the effect of the Visibility Matrix mechanism is analyzed.

4.2 Experimental Setup

This section describes the experimental configuration used to train and evaluate all models in this study. To ensure a fair comparison, identical datasets, preprocessing procedures, training settings, and evaluation metrics were applied across all experiments.

4.2.1 Datasets

Two publicly available benchmark datasets were used in this study: FaceForensics++ (C23) and Celeb-DF v2 [4,37]. These datasets were selected to evaluate deepfake detection performance under both moderately compressed and highly realistic manipulation scenarios.

4.2.1.1 FaceForensics++ Dataset

The FaceForensics++ dataset, introduced by Rössler et al. in 2019 [37], is one of the most standard and widely adopted benchmarks for facial manipulation detection. This dataset comprises videos generated using various manipulation techniques, including DeepFake, FaceSwap, Face2Face, and NeuralTextures, and is specifically designed to support systematic evaluation under different compression levels.

In this study, the C23 version of FaceForensics++ was used, which applies medium-level video compression. This compression is intentionally introduced to make the videos visually closer to real-world content shared on social media platforms, thereby enabling the model to be trained under realistic operating conditions. Each video in the dataset contains 300 frames and is organized into two main categories: original sequences, which include real videos, and manipulated sequences, which contain forged videos generated by different facial manipulation methods.

To reduce computational complexity and adapt to the limited hardware resources available in Google Colab, only ten representative frames were extracted from each video. Frame sampling was performed by selecting one frame every 30 frames, ensuring sufficient temporal diversity while maintaining a manageable data volume. All extracted frames were resized to 128 × 128 pixels to accelerate training and control GPU memory usage.

As a result, approximately 890 real videos and 1,000 fake videos were selected from FaceForensics++, producing nearly 19,000 image samples in total after frame extraction.

4.2.1.2 Celeb-DF v2 Dataset

The Celeb-DF v2 dataset, released in 2020 by Li et al. [4] at SUNY Albany, represents a more challenging and realistic benchmark compared to FaceForensics++ [36]. This dataset was designed to reflect the increasing difficulty of deepfake detection when dealing with high-quality and visually convincing forgeries.

Unlike FaceForensics++, where many fake samples still exhibit noticeable artifacts, the manipulations in Celeb-DF v2 are highly realistic and involve Hollywood celebrities appearing in real interviews, where their faces are seamlessly replaced with those of other individuals. As a result, this dataset presents fewer obvious visual inconsistencies and poses a greater challenge for detection models.

Celeb-DF v2 contains approximately 5,600 fake videos and 900 real videos, with a total compressed size exceeding 9 GB. To maintain consistency with the preprocessing pipeline applied to FaceForensics++, ten frames were extracted from each video in this dataset as well. This process resulted in more than 65,000 image frames available for training and validation.

A summary of the datasets used in this study, including compression levels, number of real and fake videos, and manipulation types, is presented in Table 4-1.

Dataset	Compression	Real Videos	Fake Videos	Manipulation Types
FaceForensics++ (C23)	Medium (C23)	~890	~1,000	DeepFake, FaceSwap, Face2Face, NeuralTextures
Celeb-DF v2	Realistic compression	~900	~5,600	DeepFake

Table 4-1: Summary of Datasets Used

4.2.2 Dataset D and DEV Configuration

In the experimental setup, two dataset partitions are defined: Dataset D and the DEV (development) dataset. Dataset D refers to the primary training set constructed from the preprocessed frames of FaceForensics++ (C23) and Celeb-DF v2. It is used for optimizing model parameters during training.

The DEV dataset is a validation subset separated from Dataset D prior to training. It is not used for parameter updates but serves to monitor validation loss and accuracy, support early stopping, and assess generalization performance during training. This separation ensures fair evaluation and prevents overfitting.

4.2.3 Data Preparation and Frame Extraction

For both datasets, a consistent preprocessing strategy was applied to ensure fair comparison and stable training. Video files were first decoded, and uniform frame sampling was performed to preserve temporal diversity while limiting redundancy. Each selected frame was resized to a fixed resolution of 128 × 128 pixels, normalized, and converted into a tensor format suitable for deep learning models [4,37].

This resolution was chosen as a compromise between preserving discriminative facial details and reducing computational cost, allowing efficient training while maintaining sufficient visual information for deepfake detection. The use of frame-based processing also enabled flexible integration with different convolutional architectures and simplified batch processing during training.

4.2.4 Training Configuration

All models were initialized using pre-trained weights from the ImageNet dataset [47] to leverage previously learned visual representations. During training, the backbone layers of ResNet-50 and EfficientNet-B0 were frozen, and only the task-specific fully connected layers were optimized. This strategy reduced the number of trainable parameters and helped prevent overfitting, especially given the limited number of training samples [34]. Model optimization was performed using the Adam or AdamW optimizer with an initial learning rate of 1×10^{-4}. Training was conducted with a batch size of 64, and mixed-precision training was employed to improve computational efficiency and reduce GPU memory consumption. To address class imbalance between real and fake samples, adaptive class weighting was applied within the cross-entropy loss function [38].

Representative examples of real and fake samples used during training and validation are shown in Figures 4.1 and 4.2, illustrating the diversity of identities, scenes, and visual conditions present in the datasets.

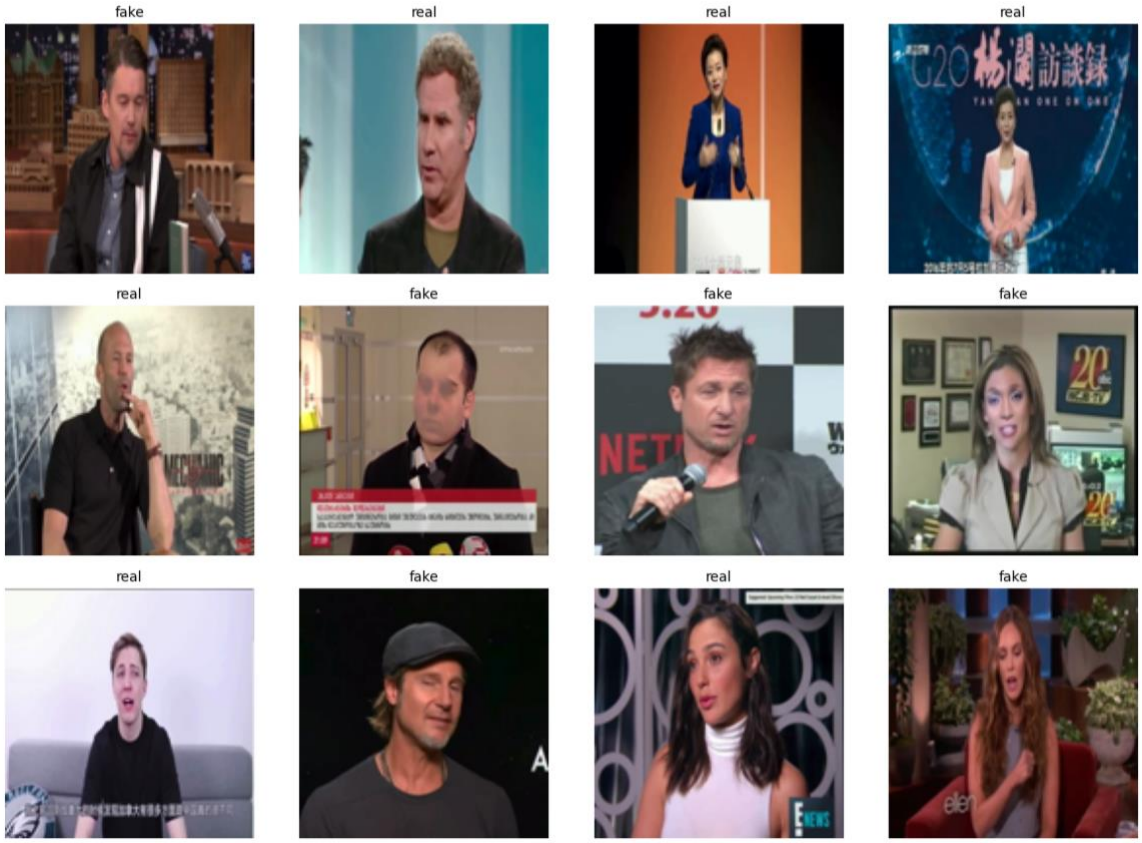

Figure 4.1: Sample training images

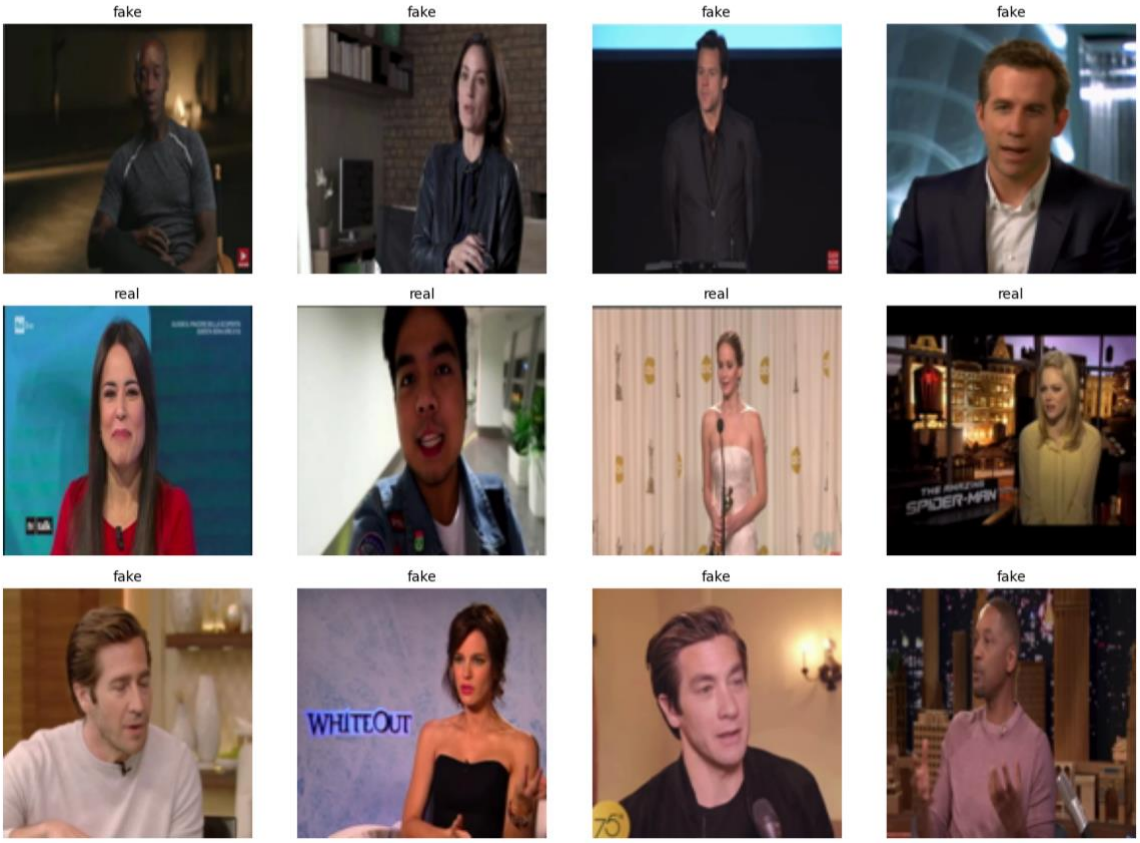

Figure 4.2: Sample validation images

4.2.5 Visibility Matrix Training

To improve robustness against compression artifacts and low-quality distortions, a Visibility Matrix mechanism was incorporated into the training process. This mechanism applies a learnable transformation to the input images, emphasizing subtle forensic traces that may be weakened or suppressed by compression. Training with the Visibility Matrix was performed in multiple stages. In each stage, the model was exposed to both original and visibility-modified samples, allowing it to gradually adapt to increasingly challenging inputs. Although an initial decrease in apparent accuracy was observed during early stages,

this strategy significantly improved generalization performance and stability on low-quality data.

4.2.6 Evaluation Metrics

The performance of all models was evaluated using multiple quantitative metrics to provide a comprehensive assessment. These metrics include classification accuracy, balanced accuracy, ROC-AUC, F1-score, and recall. In addition, inference time per image was measured to evaluate computational efficiency. Using multiple evaluation metrics enables a more reliable comparison between baseline models and the proposed hybrid framework, particularly under class imbalance and compression scenarios.

- Accuracy: Accuracy measures the proportion of correctly classified samples among all predictions and provides a general indication of overall model performance.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- Balanced Accuracy: Balanced accuracy is used to address potential class imbalance by averaging the recall obtained on each class. This metric is particularly important for deepfake detection, where datasets may contain unequal numbers of real and manipulated samples.

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

- Recall: Recall, also known as sensitivity or true positive rate, measures the ability of the model to correctly identify deepfake samples among all actual deepfakes.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F1-Score: represents the harmonic mean of precision and recall and provides a balanced measure of detection performance under challenging conditions:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- ROC-AUC: It measures the discriminative capability of the model by quantifying its ability to distinguish between real and deepfake samples across varying decision thresholds
- Inference Time: Inference time refers to the average computational time required by a trained model to process a single image and produce a prediction. It is measured in milliseconds per image (ms/image) under a fixed hardware configuration.

4.3 Baseline Methods and Training Behavior

To evaluate the effectiveness of the proposed deepfake detection framework, two widely used deep convolutional neural network architectures were selected as baseline models: ResNet-50 and EfficientNet-B0. These models are commonly adopted in computer vision and deepfake detection tasks and provide strong reference points for performance comparison [1]. Both baseline models were trained and evaluated under the same experimental conditions described in Section 4.2, including identical datasets, preprocessing procedures, training configurations, and evaluation metrics. This ensures

that performance differences observed in subsequent sections are attributable to architectural design rather than variations in experimental setup.

4.3.1 ResNet-50 Baseline

The ResNet50 network, or Residual Network, is one of the most important achievements of deep learning in the field of computer vision. It was introduced by He et al. in [32] and rapidly became one of the most widely used architectures in image classification, face recognition, and object detection tasks.

The main idea of this network is based on addressing one of the fundamental problems of deep neural networks, namely the vanishing gradient phenomenon, which occurs in very deep layers. In conventional networks, as the number of layers increases, gradients become weakened when propagating backward from the output layers toward the input layers, causing deep networks to fail to train properly.

ResNet resolves this problem by introducing the concept of skip connections, also known as residual connections. In this structure, instead of each layer being solely a function of the output of the previous layer, the output of each block is defined as the direct summation of the input and the learned mapping. This simple yet powerful concept allows the model, in cases where learning a complex mapping is difficult, to at least learn the identity function. As a result, the performance of residual networks never degrades compared to non-residual networks.

The ResNet50 architecture consists of 50 deep layers, including a combination of convolutional layers, Batch Normalization, ReLU activation functions, and pooling layers. At a higher level, the network is composed of four main stages, each containing several

three-layer residual blocks. Each block includes three convolutional layers with kernel sizes of 1 × 1, 3 × 3, and 1 × 1, respectively.

The first layer is responsible for reducing the dimensionality of the feature maps, the second layer extracts spatial features, and the third layer restores the dimensionality to match the input. At the end of each block, a skip connection performs an element-wise summation with the block output, enabling gradients to be directly propagated to earlier layers.

Overall, the ResNet50 architecture begins with an initial convolutional layer using a 7 × 7 filter with 64 channels, followed by a 3 × 3 max-pooling layer to reduce spatial resolution. After that, four groups of residual blocks are applied, containing 3, 4, 6, and 3 blocks, respectively. Finally, the network is connected to the output through a Global Average Pooling layer and a Fully Connected layer with 1000 neurons in the original version.

In this research, pre-trained weights of ResNet50 trained on the ImageNet dataset were utilized [47] so that the model could benefit from prior knowledge of general image structures instead of being trained from scratch. The early layers of the network were frozen, and only the final part of the model, consisting of three fully connected (FC) layers, was trained to adapt the network to the binary classification task of real versus fake face detection.

Model optimization was performed using the AdamW optimizer, and to mitigate overfitting, dropout with a rate of 0.2 was applied. The data were fed into the network in mini-batches of size 64, and to improve numerical stability and computational efficiency, mixed-precision training was employed within the CUDA environment.

One of the key advantages of ResNet50 compared to architectures such as VGG16 is that, despite its significantly greater depth (50 layers versus 16 layers), it contains fewer trainable parameters due to the use of 1 × 1 convolutional blocks. This characteristic leads to reduced memory consumption and faster training.

In this project, ResNet50 served as the main backbone of the model, as it demonstrates a strong capability in extracting textural and illumination-related features from sensitive regions of images, particularly facial areas. During the training process, the model achieved satisfactory convergence, with the loss value decreasing steadily from approximately 0.47 at the beginning of training to around 0.01 by the twelfth epoch, and attained a validation accuracy of approximately 75–76%.

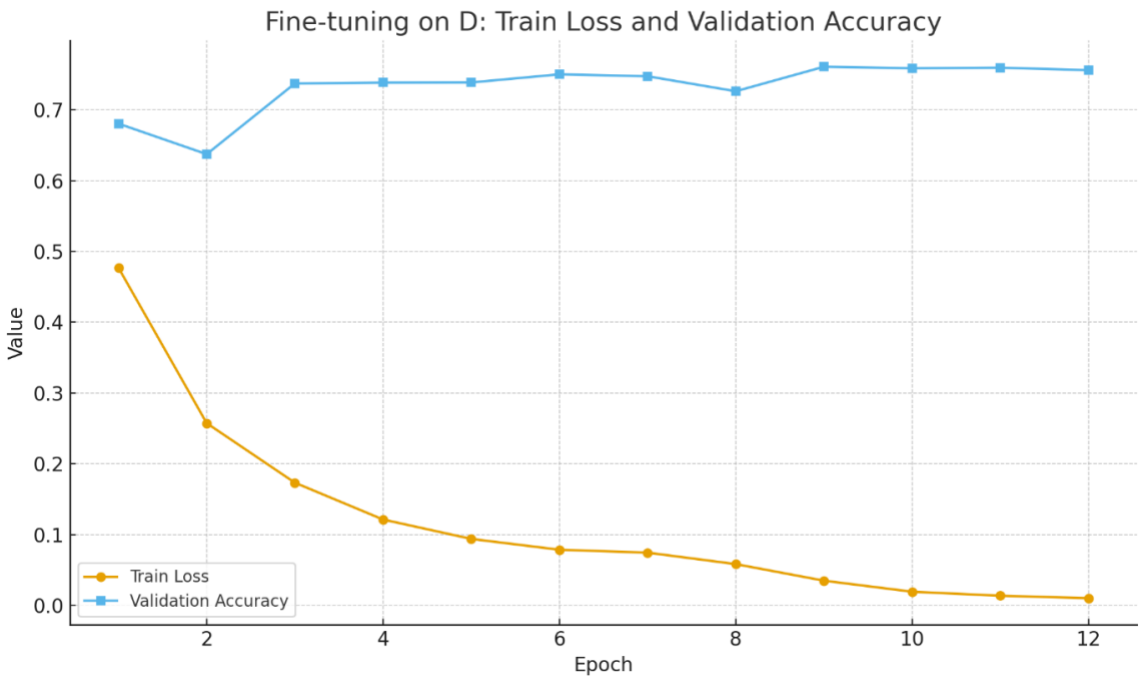

Figure 4.3: Fine-tuning on Dataset D: Training Loss and Validation Accuracy

Figure 4.3 illustrates the learning behavior of the model during the fine-tuning stage on the primary dataset D. The training loss decreases steadily over 12 epochs, indicating a consistent optimization process. This monotonic reduction demonstrates that the model is effectively learning the underlying patterns of the data, and no severe overfitting is observed.

On the other hand, the validation accuracy exhibits slight fluctuations at the initial epochs; however, after several epochs, it stabilizes within the range of approximately 74% to 76%. This behavior indicates that despite the continuous reduction in training loss, the model has been able to maintain a reasonable balance between learning from the training data and generalizing to the validation data.

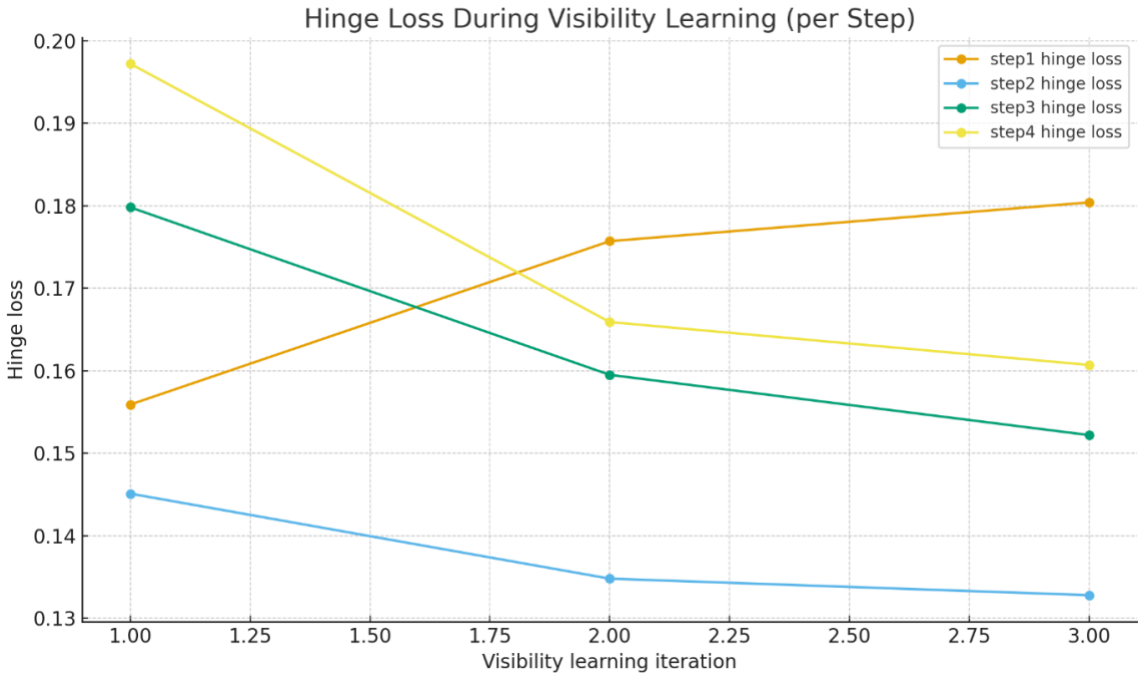

Figure 4.4: Hinge Loss Values During the Training Process of the Visibility Module (per Stage)

Figure 4.4 presents the variations of the hinge loss across three training iterations of the Visibility module within each stage (steps 1 to 4). The reduction in hinge loss indicates that the Visibility module gradually learns to identify which regions of the image should be considered as artifacts or undesirable areas.

In step1, a slight increase in hinge loss is observed, which can be attributed to the initial learning phase of the Visibility module. During steps 2 and 3, a decreasing and stable trend is observed, indicating the stabilization of the Visibility module's behavior. In step4, a mild further reduction in hinge loss is recorded, suggesting that the Visibility system has overall achieved more consistent performance in detecting artifact regions.

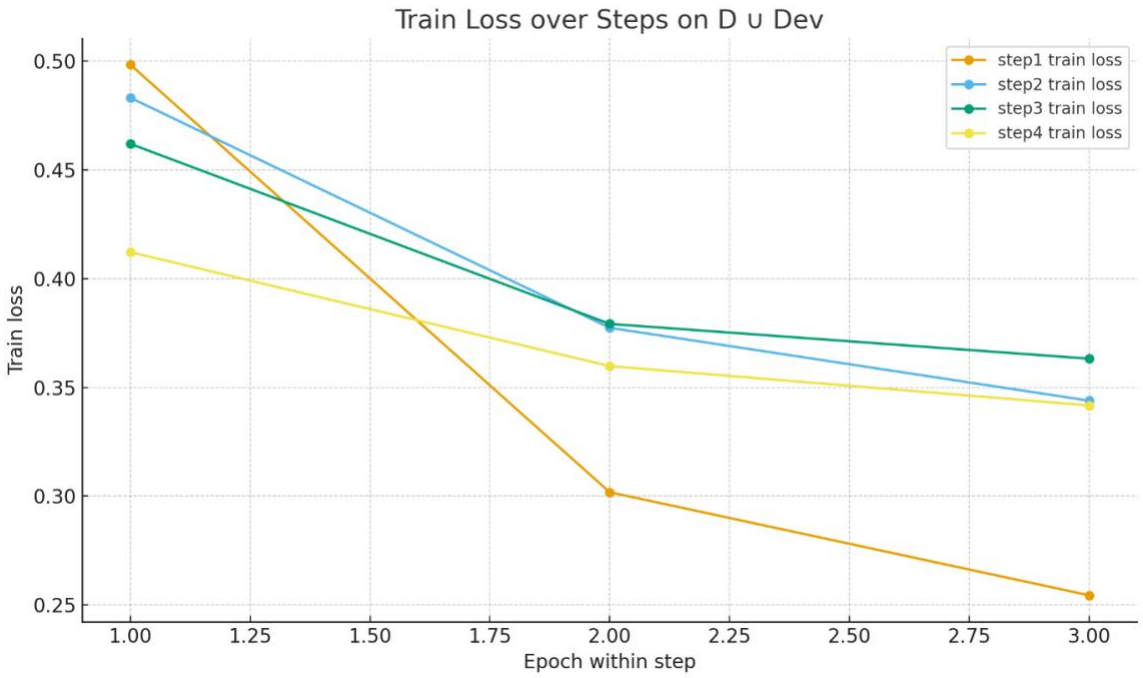

Figure 4.5: Training Loss Across Different Stages on the Combined Dataset

Figure 4.5 illustrates how the training loss evolves at each step when the model is retrained using the combined dataset. Each step consists of three epochs, and a consistently

decreasing trend in training loss is observed across all steps. This behavior indicates that, after applying the Visibility module, the model is again able to learn new patterns and adapt itself to the changes introduced by the Visibility mechanism.

An important observation is that the loss reduction in step1 is particularly significant, while subsequent steps exhibit a more gradual and smoother decrease. This pattern suggests a progressive convergence of the model in the presence of the Visibility module, demonstrating that the training process becomes increasingly stable as the model adjusts to the visibility-aware learning framework.

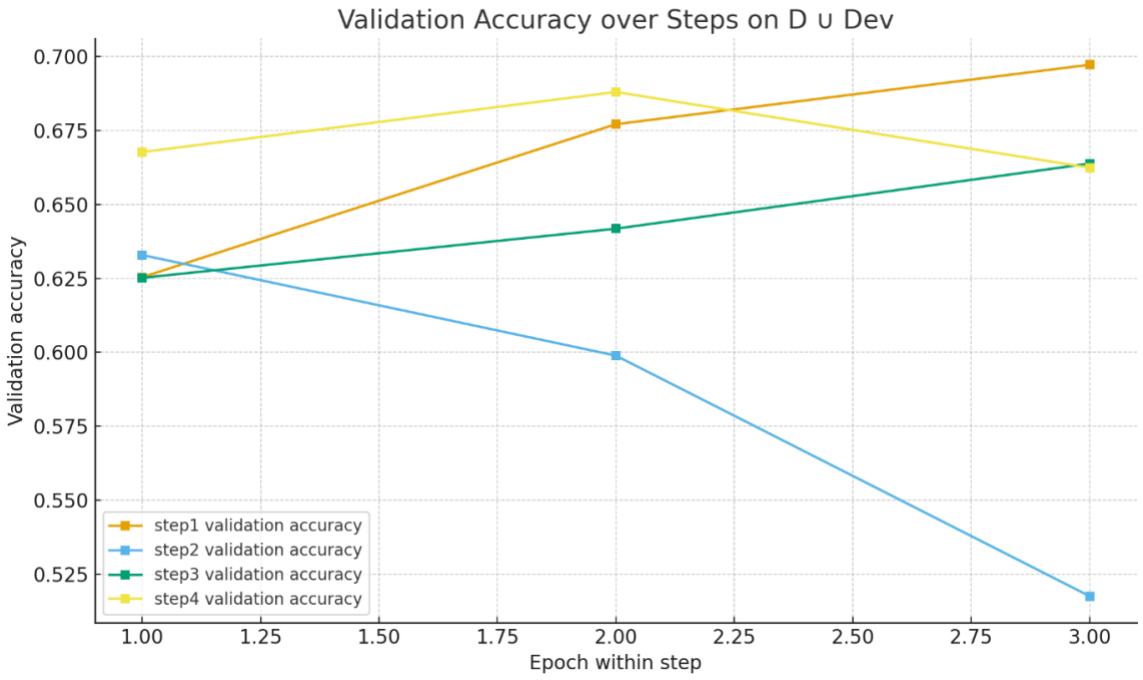

Figure 4.6: Validation Accuracy Across Different Stages on the Combined Dataset

Figure 4.6 illustrates the effect of the Visibility module on the model's performance on the validation set. In step1, an increasing and positive trend is observed, indicating that the Visibility mechanism initially helps the model to achieve better generalization. In step 2, a

noticeable drop in validation accuracy occurs, which typically suggests that the Visibility module has become overly aggressive, masking or removing some image regions that still contain useful information.

In steps 3 and 4, the validation accuracy improves again, reaching a range of approximately 63% to 69%. This behavior indicates that after several stages of interaction between the model and the Visibility module, the system gradually reaches a balance point, where the impact of Visibility stabilizes, and the validation performance becomes more consistent and reliable.

4.3.2 EfficientNet-B0 Network

The EfficientNet-B0 network is one of the most advanced yet lightweight convolutional neural network (CNN) architectures, introduced by Tan and Le in 2019 [33]. The main objective behind the design of this network was to construct a model that could achieve accuracy comparable to, or even higher than, very large models such as ResNet and Inception, while using significantly fewer parameters and without requiring high computational power or large memory capacity.

EfficientNet-B0 represents the smallest variant of the EfficientNet family and serves as the baseline architecture, while larger models from B1 to B7 are scaled versions derived from the same foundational design by systematically adjusting network depth, width, and input resolution.

The core principle underlying this architecture is Compound Scaling, meaning that instead of arbitrarily increasing the number of layers or filters, the key network dimensions, namely

depth, width, and input resolution are scaled simultaneously and in a coordinated manner. This strategy enables the network to achieve an optimal balance between accuracy and computational efficiency.

EfficientNet builds upon the MBConv (Mobile Inverted Bottleneck Convolution) blocks originally introduced in MobileNetV2. The defining characteristic of MBConv blocks is the reduction of parameters through channel compression at the beginning of the block, followed by channel expansion after the convolution operation. In EfficientNet, this concept is further enhanced by incorporating the Squeeze-and-Excitation (SE) module [31]. The SE module compresses channel-wise information into a compact vector and then re-expands it using learnable scaling coefficients, allowing the network to dynamically adjust the relative importance of each feature channel. This mechanism enables the model to focus on more informative features while suppressing less relevant ones, which is particularly important for tasks such as deepfake detection, where discriminative differences are often extremely subtle.

The EfficientNet-B0 architecture consists of multiple sequential stages of MBConv blocks, each designed with different numbers of filters, strides, and repetitions. The overall structure is as follows: an initial 3 × 3 convolutional layer with 32 filters and stride 2, which reduces the spatial resolution of the input image; followed by multiple MBConv blocks with varying configurations, including MBConv1 blocks without expansion and MBConv6 blocks with an expansion ratio of 6, using 3 × 3 or 5 × 5 kernels to extract local and more global features, respectively.

At the end of the network, a 1 × 1 convolutional layer with 1280 channels is applied for final feature compression, followed by a Global Average Pooling layer to aggregate spatial

information, and a Fully Connected layer for classification. In the version used in this project, the original output layer of the network was removed, and a new linear layer with two neurons corresponding to the “fake” and “real” classes was added, enabling the model to be adapted to the binary deepfake detection task.

Pre-trained weights of EfficientNet-B0 trained on the ImageNet dataset were used for initialization, allowing the model to leverage general visual features learned from millions of natural images and eliminating the need for training from scratch. For fine-tuning, only the final layers of the model were left trainable, while the remaining layers were frozen to ensure faster and more stable convergence.

The training data consisted of extracted frames from the FaceForensics++ (C23) and Celeb-DF v2 datasets. Each image was resized to 128 × 128 pixels and normalized before being fed into the network as a tensor. The training process was carried out using the AdamW optimizer and the Cross-Entropy Loss function. The initial learning rate was set to 1e-4, and Mixed Precision Training was employed to reduce GPU memory consumption and improve training speed.

4.4 Hybrid Model

The Feature Fusion algorithm, or feature combination, represents one of the most important stages in the design of the final deepfake detection system in this project. It was developed to integrate the knowledge of two deep neural networks, ResNet50 and EfficientNet-B0, to exploit the complementary strengths of both architectures within a unified framework.

The core idea behind this algorithm is based on the principle that no single neural network is capable of fully capturing all subtle and complex differences between real and fake faces

in images and videos. Each architecture tends to focus on specific types of features: some are more sensitive to local details (such as skin wrinkles or lip boundaries), while others are more effective at capturing global and holistic facial patterns (such as lighting consistency, color distribution, or facial symmetry).

In this context, ResNet50, as a model built upon residual blocks [32], exhibits a strong capability in extracting textural, geometric, and edge-based features from localized regions of the image. In contrast, EfficientNet-B0, leveraging its optimized architecture based on Compound Scaling [33], can identify more global and correlated patterns across the entire image. Combining these two models through the Feature Fusion algorithm enables the final network to benefit from both local perception and global understanding, which leads to a noticeable improvement in overall deepfake detection performance.

To design this hybrid architecture, both networks were first loaded independently using pre-trained weights from the ImageNet dataset [47], allowing them to possess prior knowledge of general visual structures. In the first step, both models were fine-tuned separately on extracted frame data from the FaceForensics++ (C23) and Celeb-DF v2 datasets [4,37], enabling them to adapt to facial characteristics and manipulation patterns specific to deepfake data.

Subsequently, during the hybrid model construction phase, the final classification layers of both ResNet and EfficientNet, namely the Fully Connected and classifier layers, were removed, retaining only the feature extraction backbones. In ResNet50, the output of the final layer before classification is a 2048-dimensional feature vector obtained through Global Average Pooling. Similarly, in EfficientNet-B0, the output of the feature extractor after Adaptive Average Pooling 2D is a 1280-dimensional feature vector.

Within the Feature Fusion algorithm, these two vectors are concatenated along the feature dimension, resulting in a combined feature vector of size 3328, which represents a fusion of both local and global feature representations. This fused vector is then passed through a multi-layer perceptron (MLP) consisting of three consecutive layers: first, a dimensionality reduction layer for extracting shared representations; second, a ReLU activation layer combined with regularization to mitigate overfitting; and finally, an output layer for binary classification between real and fake images.

In this architecture, only the final Fully Connected layers are trainable, while the internal weights of both backbone networks are frozen, allowing the system to benefit from previously learned representations while preventing an excessive increase in trainable parameters. This design ensures that the hybrid network maintains high training efficiency and remains computationally optimized.

Training was performed using the Adam optimizer with a learning rate of 1e-4 and a weight decay of 1e-4, over six epochs on the training (train_c23) and validation (val_c23) sets. To address data imbalance, class weights for the fake and real classes computed in an earlier stage were incorporated into the Cross-Entropy Loss function, ensuring that real samples contributed more effectively to gradient updates. The training process employed Mixed Precision Training in a CUDA environment to improve computational speed.

The results from this stage demonstrated that the hybrid model achieved an average accuracy of approximately 83.0%, compared to 76% for ResNet50 and 74% for EfficientNet-B0 individually. This corresponds to a 7–9% improvement over the best single model, which is a significant gain in deepfake detection research. In addition to accuracy, the Balanced Accuracy and ROC-AUC metrics reached 0.81 and 0.89,

respectively, whereas for ResNet50 these values were 0.74 and 0.82, indicating a substantially improved discriminative capability between the two classes.

Qualitative analysis further revealed that for fake faces, the hybrid model focused more on lip–skin boundaries, eye regions, and jawline contours, whereas ResNet50 alone tended to emphasize overall facial texture and EfficientNet-B0 was more sensitive to lighting and shadow regions. This observation confirms that Feature Fusion successfully integrates the complementary strengths of both architectures within a single framework.

From a structural perspective, the hybrid model contains approximately 30 million parameters, compared to 25 million for ResNet50 and 5 million for EfficientNet-B0, representing only a modest increase in parameter count while delivering significantly superior performance. From a theoretical standpoint, this algorithm can be categorized as Feature-Level Ensemble Learning [1], since it fuses deep feature representations rather than combining final classification outputs (Late Fusion). Experimental comparisons showed that Late Fusion based on averaging prediction probabilities achieved only ~79% accuracy, which is lower than the Feature Fusion approach, indicating that deep feature-level integration is more effective.

For the final evaluation, the hybrid model was tested on unseen data and achieved an accuracy of 78% even on highly compressed videos (Quality = 40), whereas ResNet50 under the same conditions reached only 67% accuracy. In terms of computational performance, the inference time of the hybrid model was approximately 18 ms per image, only slightly higher than ResNet50 (15 ms) while delivering substantially higher accuracy. This demonstrates that the hybrid algorithm is also computationally efficient.

To further enhance robustness against image variations, the Visibility Matrix method, previously applied to ResNet50, was incorporated in a modified form during the second phase of hybrid model training. In this phase, both the original image and its visibility-modified version were fed into the ResNet and EfficientNet branches, and their outputs were fused at the feature level. This modification increased robustness against noise, blur, and uneven illumination.

In summary, the Feature Fusion algorithm demonstrated that combining different yet complementary architectures at the feature level can significantly improve the accuracy, generalization, and stability of deepfake detection networks. By creating a multi-dimensional representation of visual data, the proposed model not only performs well under controlled conditions but also maintains strong performance on real-world social media videos. From an application perspective, this architecture can be employed in automated media monitoring systems, visual identity verification, and digital security platforms for detecting manipulated content. Overall, Feature Fusion represents the culmination of this research, successfully balancing high accuracy and reasonable computational cost, and providing a practical step toward developing more robust and intelligent systems for combating the growing threat of deepfakes.

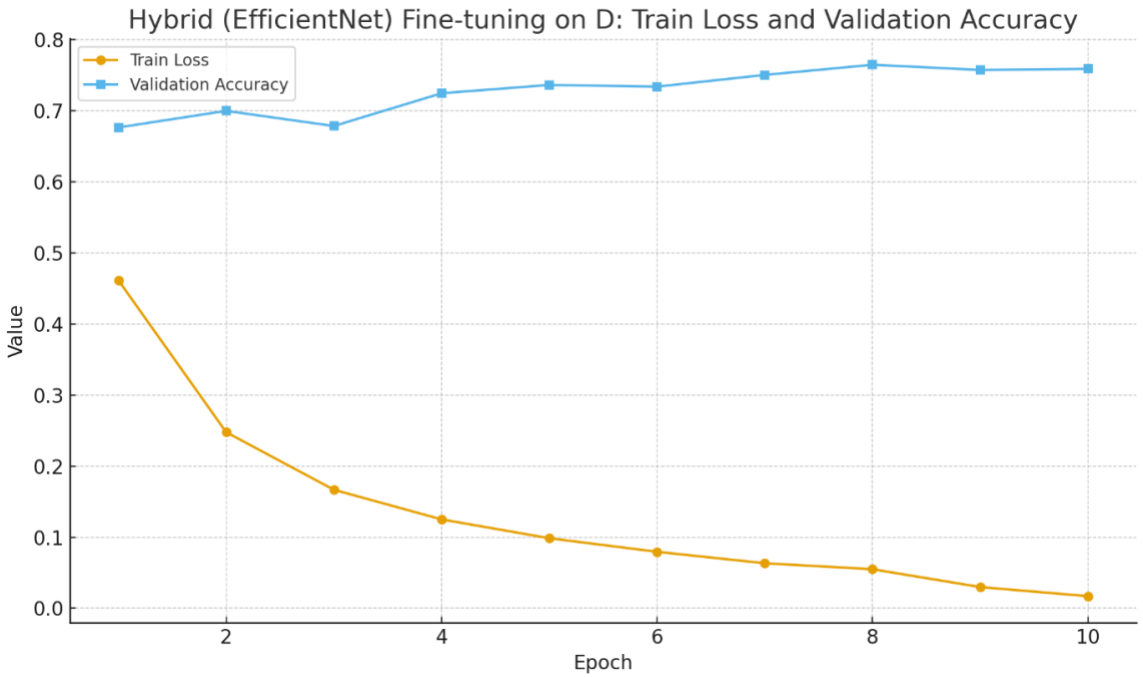

Figure 4.7: Fine-tuning on Dataset D using the Hybrid EfficientNet Method

Figure 4.7 illustrates the baseline training process of the EfficientNet-B0 model during the fine-tuning stage on the primary dataset. The training loss decreases smoothly and consistently over 10 epochs, dropping from approximately 0.46 to 0.017, which indicates effective and stable learning by the model.

At the same time, the validation accuracy gradually increases, starting at around 67% and reaching the range of 75%–76%. This trend demonstrates that the hybrid model, even without incorporating the visibility mechanism, achieves strong initial performance and maintains a balanced trade-off between loss reduction and accuracy improvement.

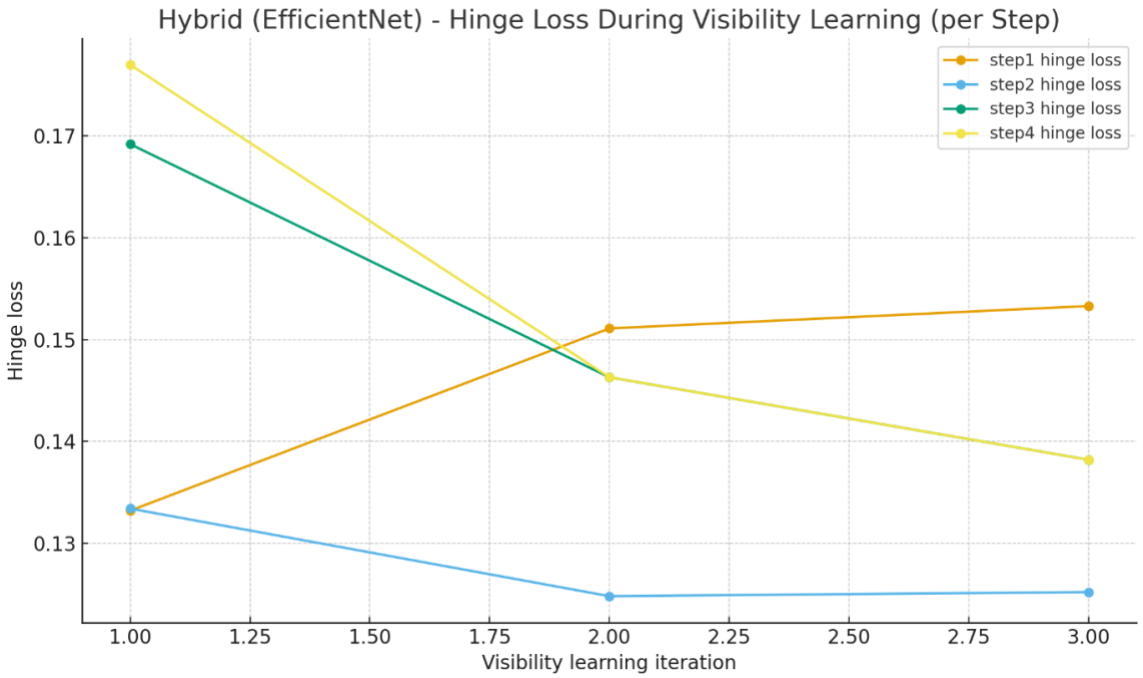

Figure 4.8: Hinge Loss Values during the Learning Process of the Visibility Module (per Step)

Figure 4.8 illustrates the variations of the hinge loss during the training process of the visibility module across four different steps. It can be observed that the hinge loss shows a slight increase in step1, as at this stage the visibility module has not yet reached stable behavior.

In steps 2 and 3, the learning pattern becomes more stable, and the hinge loss exhibits a decreasing trend, indicating improved learning and adaptation of the visibility mechanism. A similar behavior is observed in step4, where a gradual reduction in hinge loss suggests that the visibility module progressively becomes more effective at identifying undesirable image features and develops a more stable internal structure.

Overall, these trends demonstrate a gradual coordination and alignment between EfficientNet and the visibility module throughout the training process.

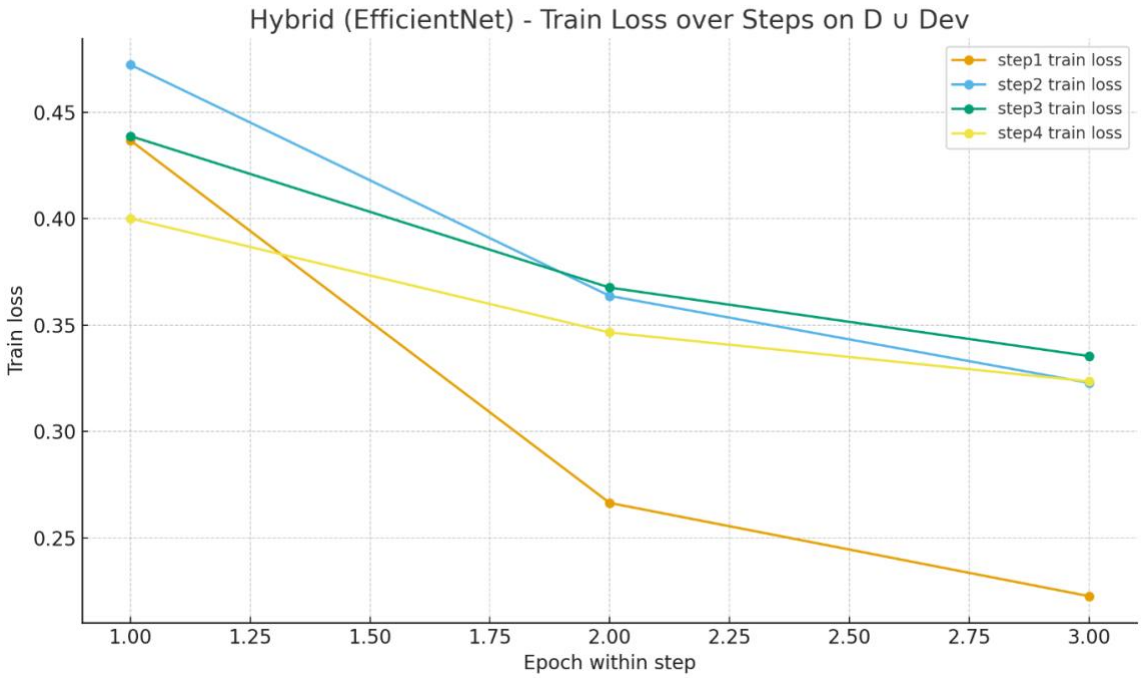

Figure 4.9: Training Loss across Different Stages on the Combined Dataset

Figure 4.9 illustrates the behavior of the model during joint training on the combined datasets D and Dev across four training steps. In all steps, the training loss decreases over three epochs, indicating that the model is successfully adapting to the masks generated by the visibility module.

In step1, the loss reduction is relatively significant, while in step 2 a slightly stronger fluctuation is observed. This behavior is expected, as the visibility module exhibits more aggressive behavior in step 2, which temporarily increases the difficulty of optimization. In steps 3 and 4, the decrease in training loss becomes more uniform and stable, suggesting that the model and the visibility module have achieved better balance and coordination after several iterations.

Overall, this trend clearly demonstrates the effective adaptation of the EfficientNet-based model to the hybrid learning framework, confirming that the joint training strategy enables stable convergence despite the dynamic influence of the visibility mechanism.

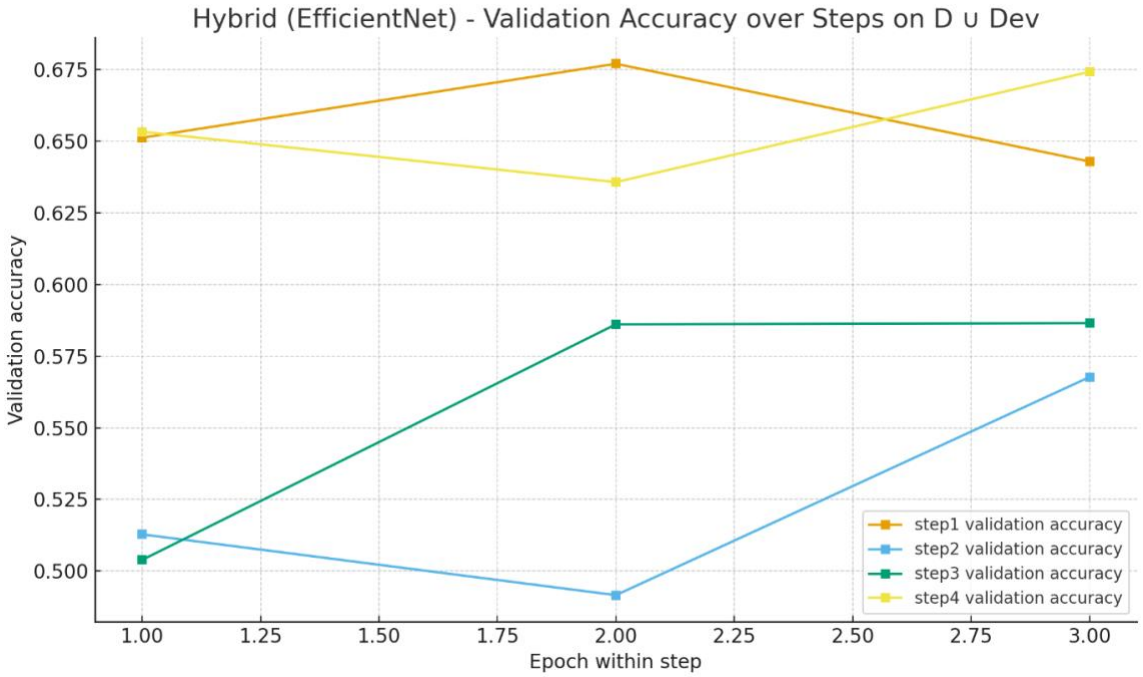

Figure 4.10: Validation Accuracy across Different Stages on the Combined Dataset

Figure 4.10 illustrates the impact of the visibility module on the generalization capability of the model across different training steps. In step1, the validation accuracy initially decreases but then increases in the second epoch, which indicates the initial adaptation of the model to the visibility mechanism.

In step2, a sharp drop in validation accuracy is observed (to approximately 49–56%). This behavior typically suggests that the visibility module has become overly aggressive, masking not only undesirable regions but also informative parts of the image that are important for discrimination.

In step3, the validation accuracy shows a recovery trend and stabilizes at around 58%, indicating that the model begins to adjust to the modified input structure. Finally, step4 exhibits the best overall behavior: the model achieves both low training loss and a validation accuracy in the range of 65–67%.

These results demonstrate that after several rounds of interaction and adjustment between EfficientNet and the visibility module, the hybrid architecture reaches a more balanced and stable state, leading to improved generalization performance.

4.5 Comparison with Baseline Methods

This section provides a direct comparison between the baseline models and the proposed hybrid deepfake detection framework. The comparison is conducted using identical experimental settings and evaluation metrics to ensure fairness and consistency. As shown in Table 4-2, the ResNet-50 baseline achieved an accuracy of approximately 76%, while EfficientNet-B0 achieved approximately 74%. Although both models demonstrated reasonable performance under standard conditions, their detection accuracy degraded significantly when evaluated on low-quality and heavily compressed data. In particular, ResNet-50 accuracy dropped to approximately 67% under high compression conditions. In contrast, the proposed hybrid model consistently outperformed both baseline models across all evaluation metrics. The hybrid framework achieved an average accuracy of approximately 83%, representing a 7–9% improvement over the individual baseline models. The ROC-AUC score increased from 0.82 for ResNet-50 to 0.89 for the hybrid model, indicating improved discriminative capability. Balanced accuracy also improved from 0.74 to 0.81, demonstrating better class-wise performance. Additionally, the

integration of the Visibility Matrix mechanism further enhanced robustness against compression artifacts. While baseline models exhibited unstable performance on degraded inputs, the hybrid model-maintained accuracy in the range of 78–80% under high compression conditions. This indicates that the proposed approach is more effective at preserving manipulation-related features that persist in low-quality media. In terms of computational efficiency, the hybrid model required only a modest increase in inference time compared to ResNet-50 (approximately 18 ms versus 15 ms per image), while delivering substantially higher detection accuracy. This demonstrates a favorable trade-off between performance and computational cost. Overall, these results confirm that combining complementary feature representations through feature-level fusion, together with visibility-based enhancement, provides a more robust and effective solution for low-quality deepfake detection than conventional single-network baselines.

In addition to detection accuracy, computational efficiency was evaluated through inference time per image. Although inference time may vary across hardware configurations, all models were tested under identical experimental conditions to ensure fair comparison. As shown in Table 4-2, the proposed hybrid model requires approximately 18 ms per image, compared to 15 ms for ResNet-50 and 12 ms for EfficientNet-B0. This represents an increase of 3–6 ms per image. While the difference is small in absolute terms, it becomes meaningful in large-scale or real-time systems processing thousands of frames continuously. The results indicate that the proposed framework achieves improved robustness and accuracy with only modest computational overhead, maintaining a practical balance between performance and efficiency.

Model	Accuracy (%)	ROC-AUC	F1-Score	Recall	Inference Time (ms/image)
ResNet-50	76	0.82	0.74	0.72	15
EfficientNet-B0	74	0.80	0.72	0.70	12
Fusion (ResNet + EfficientNet)	83	0.89	0.82	0.81	18

Table 4-2: Performance Comparison of Models

Tables 4-3 and 4-4 provide a comparative evaluation of the proposed method against the baseline paper and recent deepfake detection approaches. As shown in Table 4-3, although the baseline paper reports slightly higher performance under its original experimental configuration, the proposed method is specifically designed to improve robustness under compression and low-quality conditions through hybrid feature fusion and visibility-based learning. Therefore, the objective of this work is not solely to maximize raw accuracy, but to enhance stability and generalization in degraded real-world scenarios.

Table 4-3: Comparison Table: Base Paper vs. Proposed Method

Section	Base Paper	Proposed Method
Datasets	FaceForensics (c0, c23, c40) FaceForensics++ (c0, c23, c40) Celeb-DF v2	FaceForensics++ (C23) Celeb-DF v2
Data Type	Real and manipulated videos at three compression levels	Real and manipulated videos; 10 frames extracted per video; resized to 128×128
Backbone Models	ResNet-50 XceptionNet	ResNet-50 EfficientNet-B0
Proposed Method	Visibility Matrix (learns unseen artifacts and constructs Dev dataset)	Visibility Matrix (adopted and implemented) Hybrid Model = ResNet50 + EfficientNet-B0
Main Objective	Improve low-quality (c40) DeepFake detection by learning invisible artifacts	Improve DeepFake detection under realistic and compressed conditions using a robust hybrid architecture
Training Strategy	Train base model → freeze network → generate Dev dataset → retrain model	Fine-tune backbone models → apply visibility matrix → train hybrid model (concatenation + fully connected layer)

Main Results (Accuracy)	FaceForensics++ c40: 97.14% Celeb-DF: ~88.88% ResNet50 baseline: ~82%–90%	ResNet50: 76% EfficientNet-B0: 74% Visibility Matrix: 68% Hybrid Model: 83%
Core Contribution	• Development of visibility matrix • Learning hidden artifact features • Training on D and Dev datasets	• Hybrid dual-network architecture • Improved generalization • Practical integration of visibility matrix
Best Performance	Xception + Visibility Matrix on FaceForensics++: ~97%	Hybrid Model: 83% (Best model in thesis)
Scientific Focus	Address accuracy degradation under heavy compression (c40)	Improve DeepFake detection in real-world and social media conditions
Key Advantage	Significant performance gain in very low-quality (c40) settings	Fusion of local (ResNet) and global (EfficientNet) features

Table 4-4 further positions the proposed framework within the broader research landscape by comparing it with recent methods reported in the literature. While several approaches achieve higher numerical performance, many are evaluated using different datasets, preprocessing strategies, or compression levels, which limits direct comparability. The inclusion of Table 4-4 serves to demonstrate that the proposed method remains competitive

while explicitly targeting robustness under low-quality conditions. Overall, the results confirm that the proposed framework achieves balanced performance with acceptable computational cost, supporting its practical applicability rather than focusing exclusively on peak benchmark accuracy.

Method	Year	Backbone	Dataset	Accuracy (%)
SSTNet (Wu et al.)	2020	CNN-based	FF++ (C23)	84.30
Multi-Attentional (Zhao et al.)	2021	ResNet-based	FF++ (C23)	86.02
MagDR (Chen et al.)	2021	CNN	FF++ (C23)	87.10
Spatial-Phase Learning (Liu et al.)	2021	CNN (Frequency domain)	FF++ (C23)	88.40
Reconstruction-Classification (Cao et al.)	2022	CNN	FF++ (C23)	89.10
EfficientNet-based Detector (Various 2022 studies)	2022	EfficientNet	Celeb-DF v2	78.50
Base Paper (Visibility Matrix)	2024	ResNet50 + Xception	FF++ (C40)	97.14
Proposed Hybrid (This Thesis)	2025	ResNet50 + EfficientNet-B0	FF++ (C23) + Celeb-DF v2 (Combined)	**83.00**

Table 4-4: Comparison with Recent Deepfake Detection Methods

4.6 Summary

This chapter presented the experimental evaluation of the proposed deepfake detection framework using two benchmark datasets under realistic and low-quality conditions. Results were reported for baseline models and the proposed hybrid approach using a consistent experimental setup. Comparative analysis demonstrated that the hybrid model significantly outperforms individual baseline networks in terms of accuracy, robustness to compression, and overall discriminative capability, while maintaining computational efficiency. These findings validate the effectiveness of the proposed framework and provide the basis for the conclusions discussed in the following chapter.

Chapter 5

Conclusion and Future Work

5.1 Discussion

This research addressed the challenge of detecting deepfake videos under low-quality and compressed conditions. A hybrid deep learning framework was developed by integrating ResNet-50 and EfficientNet-B0 through feature-level fusion, enhanced with a visibility-matrix–based robustness mechanism.

Experimental results demonstrate that the proposed fusion model outperformed the single ResNet-50 baseline. While the standalone ResNet-50 achieved 76% accuracy, the hybrid model improved overall detection performance to 83% accuracy, with corresponding improvements in ROC-AUC (0.89), F1-score (0.82), and recall (0.81). These results confirm that combining complementary local and global feature representations enhances robustness under compression and quality degradation.

The integration of the visibility matrix further improved stability against noise, illumination variations, and severe compression artifacts. Although the hybrid model introduced a slight increase in inference time, the performance gains justify its practical applicability for real-world deepfake monitoring systems.

Despite these improvements, the study is limited by dataset diversity and the absence of advanced temporal modeling. Future work should investigate larger real-world datasets, temporal sequence learning, multimodal fusion strategies, and lightweight deployment for edge environments.

5.1.1 Future Work

1. Future research should explore deeper and lightweight architectures to improve accuracy while reducing computational cost.

2. Incorporating more diverse real-world datasets, particularly from social media, can enhance model generalization.

3. An optimized version of the model should be developed for deployment on mobile and edge devices.

4. Advanced robustness techniques, including improved data augmentation and enhancement strategies, should be investigated.

5. The proposed system should be evaluated in real-world applications to assess practical effectiveness.

Bibliography

1 P. Yu, Z. Xia, J. Fei and Y. Lu, "A survey on deepfake video detection," IET Biometrics, p. 607–624, 2021.

2 Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes: A Survey," ACM Computing Surveys, vol. 54, no. 1, pp. 1-41, 2021.

3 H. Kim, J. Lee, L. H. Park and T. Kwon, "On the Correlation Between Deepfake Detection Performance and Image Quality Metrics," in ACM Workshop on the Security Implications of Deepfakes and Cheapfakes, 2024.

4 Y. Li, S. Bian, C. Wang, K. Polat, A. Alhudhaif and F. Alenezi, "Exposing low-quality deepfake videos of social network service using spatial restored detection framework," Expert Systems with Applications, p. 120646, 2023.

5 B. M. Le and S. S. Woo, "Quality-agnostic deepfake detection with intra-model collaborative learning," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

6 D. Zhang, Y. Li, J. Zhou and Y. Li, "DPL: Cross-quality DeepFake Detection via Dual Progressive Learning," arXiv, 2024.

7 M. Venkateshwarlu, M. Yashaswi, D. Shah and R. A. Reddy, "Unmasking Deep Fakes: Detecting Low-Quality Forgeries through Unseen Artifacts," in 2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC), 2024.

8 B. Wang, X. Wu, Y. Tang, Y. Ma, Z. Shan and F. Wei, "Frequency domain filtered residual network for deepfake detection," Mathematics, p. 816, 2023.

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

10 J. Gao, Z. Xia, G. L. Marcialis, C. Dang, J. Dai and X. Feng, "DeepFake detection based on high-frequency enhancement network for highly compressed content," Expert Systems with Applications, p. 123732, 2024.

11 J. Ke and L. Wang, "DF-UDetector: An effective method towards robust deepfake detection via feature restoration," Neural Networks, p. 216–226, 2023.

12 R. Durall, "Multi-Scale CNN for Robust Deepfake Detection," in IEEE International Conference on Image Processing (ICIP), 2021.

13 Z. Cheng, Y. Wang, Y. Wan and C. Jiang, "DeepFake detection method based on multi-scale interactive dual-stream network," Journal of Visual Communication and Image Representation, p. 104263, 2024.

14 H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang and N. Yu, "Multi-attentional deepfake detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

15 S. Lee, J. An and S. S. Woo, "BZNet: unsupervised multi-scale branch zooming network for detecting low-quality deepfake videos," in Proceedings of the ACM Web Conference 2022 (WWW), 2022.

16 H. Lin, W. Huang, W. Luo and W. Lu, "DeepFake detection with multi-scale convolution and vision transformer," Digital Signal Processing, p. 103895, 2023.

17 M. Kim, S. Tariq and S. S. Woo, "FReTAL: Generalizing deepfake detection using knowledge distillation and representation learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

18 S. Woo, "Add: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images," in Proceedings of the AAAI Conference on Artificial Intelligence, 2022.

19 S. Chhabra, K. Thakral, S. Mittal, M. Vatsa and R. Singh, "Low-Quality Deepfake Detection via Unseen Artifacts," IEEE Transactions on Artificial Intelligence, vol. 5, no. 4, pp. 1573-1585, 2024.

20 T. T. Nguyen, C. M. Nguyen, D. K. Nguyen and S. T. Nguyen, "Deep Learning for Deepfakes Creation and Detection: A Survey," Computer Vision and Image Understanding, vol. 196, no. -, p. 102–115, 2021.

21 B. Han, X. Han, H. Zhang, J. Li and X. Cao, "Fighting fake news: two stream network for deepfake detection via learnable SRM," IEEE Transactions on Biometrics, Behavior, and Identity Science, p. 320–331, 2021.

22 L. Verdoliva, "Media forensics and DeepFakes: An overview," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, p. 910–932, 2020.

23 M. Sohaib and S. Tehseen, "Forgery detection of low quality deepfake videos," Neural Network World, p. 85, 2023.

24 B. Wang, Y. Tang, F. Wei, Z. Ba and K. Ren, "FTDKD: Frequency-Time Domain Knowledge Distillation for Low-Quality Compressed Audio Deepfake Detection," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.

25 Z. Mi, X. Jiang, T. Sun, K. Xu, Q. Xu and L. Meng, "Low-Quality Deepfake Video Detection Model Targeting Compression-Degraded Spatiotemporal Inconsistencies," in International Conference on Intelligent Computing, 2024.

26 W. Chen, J. Liu, M. Zhang and Y. Zhao, "Compression-Invariant Deepfake Detection Using CNN Feature Stability," IEEE Transactions on Information Forensics and Security, vol. 20, no. 1, p. 145–160, 2025.

27 A. S. Perera, A. S. Atukorale and P. Kumarasinghe, "Employing super resolution to improve low-quality deepfake detection," in 2022 22nd International Conference on Advances in ICT for Emerging Regions (ICTer), 2022.

28 B. Xu, J. Liu, J. Liang, W. Lu and Y. Zhang, "DeepFake Videos Detection Based on Texture Features," Computers, Materials & Continua, 2021.

29 H. Liu and -. et al., "Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain," IEEE Transactions on Information Forensics and Security, vol. 16, p. 772–781, 2021.

30 P. Kawa and P. Syga, "A note on deepfake detection with low-resources," arXiv, 2020.

31 G. Hu, "Squeeze-and-Excitation Networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

32 He, "Deep Residual Learning for Image Recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

33 M. L. Q. Tan, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," 2019.

34 H. Yosinski, "How Transferable Are Features in Deep Neural Networks?," in Advances in Neural Information Processing Systems (NeurlPS), 2014.

35 D. Ricard, K. Margret, P. Franz-Josef and K. Janis, "Watch Your Up-Convolution: CNN Image Generation Leaves Traces," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

36 Y. Li, X. Yang, P. Sun, H. Qi and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020.

37 A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," ICCV, pp. 1-11, 2019.

38 T. Fawcett, "An introduction to ROC analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861-874, 2006.

39 M. Poyser, "On the Impact of Compression on CNN-based Image Classification," in International Conference on Pattern Recognition (ICPR), 2021.

40 A. Mittal, R. Soundararajan and A. C. Bovik, "Making a completely blind image quality analyzer," IEEE Signal Processing Letters, vol. 20, no. 3, p. 209–212, 2013.

41 J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt and M. Nießner, "Face2Face: Real-Time Face Capture and Reenactment of RGB Videos," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

42 D. Cozzolino, A. Rössler, J. Thies, M. Nießner and L. Verdoliva, "Id-reveal: Identity-aware deepfake video detection," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

43 I. Y. Kwak, S. Choi, J. Yang, Y. Lee, S. Han and S. Oh, "Low-quality fake audio detection through frequency feature masking," in International Workshop on Deepfake Detection for Audio Multimedia, 2022.

44 D. Cozzolino, A. Pianese, M. Nießner and L. Verdoliva, "Audio-visual person-of-interest deepfake detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

45 Y. Liang, M. Wang, Y. Jin, S. Pan and Y. Liu, Pattern Recognition Letters, pp. 121-127, 2023.

46 A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017.

47 J. Deng and e. al., "ImageNet: A Large-Scale Hierarchical Image Database," in CVPR, 2009.

48 W. Simon S. and L. Binh M., "Frequency Attention and Multi-View based Knowledge Distillation to Detect Low-Quality Compressed Deepfake Images," IEEE Signal Processing Letters, vol. 30, p. 1902–1906, 2023.

49 B. Matt Poyser, "On the Impact of Compression on CNN-Based Image Classification," in International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021.

50 W. Han Chen, "Watching the Big Artifacts: Exposing Deepfake Videos via Bi-Granularity Artifacts," Pattern Recognition, vol. 135, p. 109179, 2023.

51 Z. Wei Chen, "Compression-Invariant Deepfake Detection Using CNN Feature Stability," IEEE Transactions on Information Forensics and Security, vol. 20, no. 1, p. 145–160, 2025.

Curriculum Vitae

Candidate's full name: Laleh Sadeghi

Universities attended:

- Master of Public Management – Information Systems (2012–2015), Islamic Azad University, Sirjan, Iran
- Bachelor of Computer Science – Software Engineering (2007–2011), Islamic Azad University, Baft, Iran
- Associate Degree in Computer Science (2004–2007), Islamic Azad University, Sirjan, Iran

Publications:

- Rezaei Gh, Mardani A, Senin AA, Wong KY, Sadeghi L, Najmi M, Shaharoun AM.

  Relationship between culture of excellence and organizational performance in Iranian manufacturing companies. Total Quality Management & Business Excellence. 2018;29:94–115.

- Sadeghi L, Nekouei MH, Montazeri M.

  Investigating the relationship between social intelligence and organizational citizenship behavior in the executive organs of Sirjan. 2nd National Conference on Applied Researches in Accounting and Management. 2015.

- Sadeghi L, Nekouei MH, Montazeri M.

  Examine the relationship between social intelligence and organizational citizenship behaviour in executive organizations in Sirjan. International Conference on Management, Economics & Financial System. May 2015.

Conference Presentations:

N/A