

ATTENTION NETWORK FOR DEEP FAKE DETECTION

A PROJECT REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
AWARD OF THE DEGREE
OF

**MASTER OF TECHNOLOGY
IN
COMPUTER SCIENCE**

Submitted By:
SNEHA KUMARI
(2K22/CSE/30)

Under the supervision of
Mr. Kavinder Singh



DEPARTMENT OF COMPUTER ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly, Delhi College of Engineering)

Bawana Road, Delhi-110042

MAY, 2024

CANDIDATE'S DECLARATION

I, **Sneha Kumari, 2K22/CSE/30** student of M. Tech., Computer Science, hereby declares that the project Dissertation titled “**Attention Network for Deep Fake Detection**” which is submitted by me to the Department of Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship, or other similar title or recognition.

Place: Delhi

Date: 30th May 2024

Sneha Kumari

(2K22/CSE/30)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly, Delhi College of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled “Attention Network for Deep Fake Detection” which is submitted by Sneha Kumari, 2K22/CSE/30 from the Department of Computer Science and Engineering, Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the Degree of Master of Technology, is a record of the project work carried out by the students under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi
Date: 30th May 2024

Mr. Kavinder Singh
SUPERVISOR
(Assistant Professor)
Dept. of Computer Science and Engg.

ABSTRACT

The swift progression of machine learning techniques has led to the creation of "deepfakes," hyper-realistic synthetic media generated using algorithms like generative adversarial networks (GANs) and autoencoders. While the deepfake technology has beneficial applications in areas like entertainment and education, it poses serious significant risks such as misinformation, identity theft, and reputational damage. Therefore, the development of robust detection mechanisms, particularly those leveraging attention networks, is of paramount importance.

This thesis focuses on the importance of attention networks for detecting deepfake, offering a comprehensive analysis of their effectiveness and challenges compared to traditional and contemporary methods. Attention networks, which enhance detection by focusing on critical regions of an image, are evaluated alongside convolutional neural networks (CNNs), multimodal detection techniques, adversarial training, transformers, and frequency-based models. Performance metrics like accuracy and the area under the receiver operating characteristic curve (AUC-ROC) are used to assessing models. The thesis emphasizes the significance of temporal coherence in video analysis and the role of frequency filters in identifying subtle artifacts. Attention-based methods are shown to offer superior performance in detecting fine-grained manipulations, achieving high accuracy and AUC-ROC scores. However, these models also face challenges related to computational complexity and generalization to novel deepfake techniques.

The findings underscore the potential of attention networks to enhance deepfake detection, particularly in the real-world applications like social media moderation, news verification, and cybersecurity. This research not only helps advance but also helps in the understanding of deepfake detection, also bridges the gap between academic innovation and practical implementation. Future directions for research are suggested, focusing on improving computational efficiency, robustness, and ethical deployment of these technologies.

By exploring the capabilities and limitations of attention networks in deepfake detection, this thesis helps continue the ongoing efforts to protect the integrity of digital content in an increasingly complex digital landscape.

ACKNOWLEDGEMENT

We would like to express our heartfelt gratitude to all the individuals who have supported and assisted us throughout our M. Tech Major Project. First and foremost, we would like to thank our project supervisor, “**Mr. Kavinder Singh**”, *Assistant Professor*, Department of Computer Science and Engineering, Delhi Technological University for his constant guidance, support, and encouragement throughout the project. We are indebted to him for sharing his knowledge, expertise, and valuable feedback that helped us in shaping the project.

We would like to extend our sincere thanks to the **Vice Chancellor** of Delhi Technological University, “**Prof. Prof. Prateek Sharma**” and the faculty members of the Department of Computer Engineering for their support and encouragement throughout our academic journey. We are grateful to “**Dr. Vinod Kumar**”, *Head of Department*, Department of Computer Engineering, Delhi Technological University, Delhi for his valuable suggestions and feedback on our project work.

We are also thankful to our parents for their constant support and motivation, which has been our driving force throughout our academic pursuits. We would like to express our gratitude to our friends and classmates who have been a constant source of inspiration and motivation.

Finally, I would like to thank all the participants who participated in the study, without whom this research would not have been possible. I express my sincere gratitude to all the individuals who have directly or indirectly contributed to the success of my project

Sneha Kumari

(2K22/CSE/30)

TABLE OF CONTENTS

| | |
|--|-------------|
| Declaration | i |
| Certificate | ii |
| Abstract | iii |
| Acknowledgment | v |
| Contents | vi |
| List of Tables | viii |
| List of Figures | ix |
| List of Symbols, Abbreviations | x |
| | |
| 1 INTRODUCTION..... | 1 |
| 1.1 Context..... | 1 |
| 1.2 Applications and Misuses..... | 2 |
| 1.3 Significance of Deepfake Detection..... | 2 |
| 1.4 Research Objectives..... | 3 |
| 1.5 Scope of the Study..... | 4 |
| 1.6 Background of Deepfake Detection..... | 4 |
| 1.7 Uses of deepfakes..... | 5 |
| 1.8 Components of deepfake detection..... | 5 |
| 1.9 Latest Advances in deepfake detection..... | 6 |
| | |
| 2 LITERATURE REVIEW..... | 8 |

| | |
|---|-----------|
| 2.1 Deepfake Detection Methods..... | 8 |
| 2.2 Methods based on attention networks..... | 11 |
| 3 METHODOLOGY..... | 17 |
| 3.1 Dataset Used and Preprocessing..... | 17 |
| 3.2 Evaluation Metrics..... | 17 |
| 3.3 Proposed Model Architecture..... | 18 |
| 3.4 Architecture Description..... | 18 |
| 3.5 Model Training..... | 19 |
| 3.6 Experimental Setup..... | 20 |
| 3.7 Results of proposed architecture..... | 20 |
| 4 EXPERIMENTAL RESULTS AND ANALYSIS..... | 21 |
| 4.1 Comparison of study results..... | 21 |
| 4.2 Comparative Analysis of Deepfake Detection Models:..... | 25 |
| 4.3 Performance Comparison: Attention..... | 29 |
| 4.4 Results | 31 |
| 4.5 Discussion on models of the study..... | 32 |
| 4.6 Analysis on Proposed Architecture..... | 33 |
| 5 CONCLUSION..... | 37 |
| 6 REFERENCES..... | 38 |

LIST OF TABLES

| Sr. No. | Tables | Pg. No. |
|----------------|---|----------------|
| 3.7 | Comparison of our model with other models | 20 |
| 4.1 | Comparison of Methods | 26 |
| 4.4.1 | Training and testing on same dataset | 32 |
| 4.4.2 | Cross dataset testing | 32 |

LIST OF FIGURES

| Sr. No. | Figure | Page No. |
|----------------|-----------------------------|-----------------|
| 3.1 | Proposed Model Architecture | 26 |

LIST OF ABBREVIATIONS

| Abbreviation | Full Form |
|---------------------|--|
| CNN | Convolutional Neural Networks |
| GAN | Generative Adversarial Networks |
| A.I. | Artificial Intelligence |
| RNN | Recurrent Neural Network |
| AUC-ROC | Area under the Receiver operating characteristic curve |
| LSTM | Long Short-Term Memory |
| LQ | Low Quality |
| MQ | Medium Quality |
| HQ | High Quality |
| CMF | Cross Modality Fusion |
| ViViT | Video Vision Transformer |
| FF++ | Face Forensics |
| DFDC | Deepfake Detection Challenge |

CHAPTER-1: INTRODUCTION

1.1 Context

The advent of advanced machine learning techniques has revolutionized various fields, including image and video manipulation. One of the most prominent outcomes of this technological advancement is the creation of "deepfakes," a term derived from "deep learning" and "fake." Deepfakes refer to fake media in which a person in real image with someone else's likeness, often with stunning realism. This technology leverages deep learning algorithms, by exclusively using generative adversarial networks (GANs) and autoencoders, making it difficult to differentiate between real and manipulated content.

The term "deepfake" was first popularized in late 2017 when a Reddit user started posting doctored videos, swapping celebrities' faces with those of pornographic actors. This marked the beginning of widespread awareness and concern regarding the potential misuse of AI-driven media manipulation. The realistic nature of these deepfakes posed a significant challenge, as traditional forensic techniques and human perception struggled to detect the artificial alterations.

1.1.1 Emergence of Generative Models

Deepfake technology primarily relies on generative models, with GANs being one of the most influential. Introduced by Ian Goodfellow et al. and his colleagues, 2014. The model is made up of two neural nets: a generator and a discriminator. The generator helps create fake images, while the discriminator attempts to discriminate between real and synthetic images. Through iterative training, the generator becomes adept at producing highly realistic images that can deceive the discriminator, and by extension, human observers.

Autoencoders, another cornerstone of deepfake technology, are NN used to learn efficient codings of input data. Deep Fakes are often employed to encode the features of a person's face and then decode them onto another person's face, facilitating realistic facial swapping and manipulation.

1.2 Applications and Misuses

The applications of deepfake technology are diverse and span several industries. In the entertainment industry, deepfakes have been used to create special effects, resurrect deceased actors, or produce more realistic dubbing and translations. This technology has also found use in education, providing realistic historical reenactments and training simulations. In marketing, companies employ deepfakes to create engaging advertisements or personalized content for consumers.

Despite these benign applications, deepfake technology has garnered notoriety for its potential for misuse. Deepfakes could be weaponized to spread dangerous misinformation and disinformation, creating fake news that can manipulate public opinion and further undermine trust in the media. They pose risks in cybersecurity, as they can be used for identity theft, fraud, and unauthorized access to secure systems. Additionally, deepfakes have been used for creating non-consensual explicit content, leading to significant reputational damage and harassment.

1.3 Significance of Deepfake Detection

The rise of deepfakes presents significant challenges across multiple domains, including politics, entertainment, and personal privacy. Deepfakes could be used to spread misinformation, create malicious hoaxes, and damage reputations. The potential for deepfakes to be employed in cybercrime and misinformation campaigns has prompted growing concern among governments, technology companies, and the general public. Therefore, developing robust deepfake detection methods is important in preserving the integrity of digital media and further help protecting individuals and organizations from the adverse effects of such deceptive practices.

The rise of deepfake technology has led to significant developments in artificial intelligence, digital media manipulation, and cybersecurity, necessitating robust detection mechanisms to counter its malicious uses. The ability of latest models to create highly realistic fake media is a threat to the integrity of digital media, prompting the need for advanced detection techniques.

The challenges associated with detecting deepfakes are multifaceted. Traditional forensic methods, which rely on identifying inconsistencies in physical and geometric properties of images and videos, are often inadequate against sophisticated deepfakes that minimize detectable artifacts. Additionally, as deepfake algorithms continue to improve, the gap between genuine and fake content becomes increasingly difficult to bridge.

Machine learning and AI-based detection methods have emerged as critical tools in this battle. These methods leverage deep learning models to analyze and classify media content, identifying subtle artifacts and inconsistencies that may indicate manipulation. However, the rapid evolution of deepfake technology requires continuous advancements in detection techniques to stay ahead of new and emerging threats.

1.3.1 The Growing Importance of Deepfake Detection

Given the potential impact of deepfakes on various sectors, the importance of developing robust detection mechanisms cannot be overstated. Governments, technology companies, and researchers are increasingly investing in deepfake detection to safeguard the integrity of digital media. Initiatives such as the DeepFake Detection Challenge and the Partnership on AI have been instrumental in advancing the field by providing comprehensive datasets and fostering collaboration among stakeholders.

The detection of deepfakes is not only a technical challenge but also a societal one. Ensuring the authenticity of digital media is crucial for maintaining public trust in media, protecting individual privacy, and preventing the misuse of AI technologies. As deepfake tech. continues to evolve, the development of effective detection methods will remain a critical area of research and innovation, essential for mitigating the risks associated with this powerful yet potentially dangerous technology.

1.4 Research Objectives

This thesis aims to investigate and develop effective methods for detecting deepfakes. The primary objectives are:

- To understand the underlying technologies used to create deepfakes.
- To identify the challenges and limitations of current deepfake detection techniques.
- To explore and develop advanced detection algorithms that can improve accuracy and robustness.
- To evaluate the performance of these algorithms on benchmark datasets.

1.5 Scope of the Study

This study focuses on the technical aspects of deepfake detection, including feature extraction, machine learning models, forensic analysis, and hybrid approaches. It also explores the latest advancements in the field, such as improved machine learning models, multimodal detection, adversarial training, explainable AI, and real-time detection capabilities. The research is grounded in both theoretical analysis and practical implementation, aiming to bridge the gap between academic research and real-world application.

1.6 Background of Deepfake Detection

1.6.1 What is a Deepfake

Deepfakes are hyper-realistic digital forgeries created using advanced machine learning techniques, particularly GANs and deep learning algorithms. These technologies enable the creation of synthetic media in which the likeness of one person is replaced with another in a convincing manner. The rise of deepfake technology has led to significant developments in artificial intelligence, digital media manipulation, and cybersecurity.

1.6.2 How Deepfakes are Created

The creation of deepfakes involves training neural networks on large datasets of video and audio recordings of target individuals. The most common technique employed is the use of GANs, where two neural nets—the generator and the discriminator—compete against each other. The generator generates fake content, while the discriminator is used for classification. Through iterative training, the generator improves its output until the fake content becomes indistinguishable from real content to the discriminator. This adversarial process results in highly realistic synthetic media. Another approach involves autoencoders, where the model learns to encode and decode input data. By training autoencoders on a person's face, it becomes possible to manipulate and alter the facial features to match another person's expressions and movements.

1.7 Uses of Deepfakes

Deepfakes have various applications, both benign and malicious:

Entertainment and Media: Deepfakes are used in the entertainment industry to create special effects, resurrect deceased actors, or produce realistic dubbing and translations.

Education and Training: They can be employed for educational purposes, such as creating historical reenactments or generating realistic training simulations.

Advertising and Marketing: Companies use deepfakes to create engaging advertisements or to personalize marketing content for individual consumers.

However, the malicious uses of deepfakes have raised significant concerns:

Misinformation and Disinformation: Deepfakes are used to spread false information, synthesize fake news, and manipulate public opinion.

Fraud and Identity Theft: They can be exploited for financial fraud, such as impersonating individuals to gain unauthorized access to secure systems or commit fraud.

Reputation Damage and Harassment: Deepfakes can be misused to create a non-consensual explicit content, damaging the reputations of individuals and leading to harassment.

1.8 Components of Deepfake Detection

The detection of deepfakes involves various techniques and components:

Feature Extraction: Identifying unique features or artifacts in media that may indicate manipulation. This includes inconsistencies in lighting, shadows, and reflections, as well as anomalies in facial movements and audio signals.

Machine Learning Models: Utilizing machine learning algorithms to analyze and classify media content. Convolutional neural networks (CNNs) and recurrent neural

networks (RNNs) are usually used to detect spatial and temporal inconsistencies in videos.

Forensic Analysis: Applying traditional digital forensics techniques to examine the physical and geometric properties of media files. This includes analyzing metadata, compression artifacts, and noise patterns.

Hybrid Approaches: Combining deep learning and forensic techniques to enhance the accuracy of detection systems.

Benchmark Datasets: Using standardized datasets for training and evaluating detection models. Popular datasets include FaceForensics++, DeepFake Detection Challenge Dataset, and Celeb-DF.

1.9 Latest Advances in Deepfake Detection

The field of deepfake detection is rapidly evolving, with continuous advancements aimed at improving the accuracy of detection methods. Some of the latest advances include:

Improved Machine Learning Models: Recent developments in deep learning have led to more sophisticated models capable of detecting subtle artifacts in deepfake videos. Techniques such as attention mechanisms and transformers are being incorporated into detection models to enhance their performance.

Multimodal Detection: Researchers are exploring integration of many modalities like visual, audio, and textual information—to improve detection accuracy. Multimodal approaches can leverage inconsistencies across different data types to identify deepfakes more effectively.

Adversarial Training: To counter adversarial techniques used by deepfake creators, detection models are being trained with adversarial examples. This involves exposing the models to manipulated data during training, improving their ability to detect tampered content in real-world scenarios.

Explainable AI: Efforts are being put to develop explainable AI techniques for deepfake detection, enabling the models to provide interpretable and transparent results. This helps in understanding the decision-making process of models and to build trust in their predictions.

Real-Time Detection: Advances in computational efficiency are enabling real-time detection of deepfakes. This is particularly important for applications requiring immediate verification, such as live video streams and social media content moderation.

Collaborative Initiatives: Organizations and research institutions are collaborating to create comprehensive datasets, share knowledge, and develop standardized benchmarks for deepfake detection. Initiatives like the DeepFake Detection Challenge and the Partnership on AI are driving progress in this field.

Chapter 2 - Literature Review

Deepfake detection has witnessed significant advancements over the years, transitioning from basic forensic techniques to sophisticated machine learning models that leverage large datasets and advanced architectures. This review covers the evolution of these techniques, describing the methodologies, models, performance metrics, and limitations.

2.1 Deepfake Detection Methods

Forensic Techniques: Early methods for deepfake detection were focused on forensic analysis, which involved identifying inconsistencies in the physical and geometric properties of images and videos, such as lighting, shadows, and reflections. These methods relied on detecting artifacts like noise patterns and compression inconsistencies. However, these forensic-based models often struggled with high-quality deepfakes that minimized detectable artifacts. Performance metrics such as accuracy and AUC were not uniformly reported, making it difficult to assess their effectiveness comprehensively.

FaceForensics++ and XceptionNet: A major development in deepfake detection was the introduction of the FaceForensics++ dataset presented by Rössler et al. [1]. This dataset provided a comprehensive benchmark for training and evaluating detection models. The researchers employed Convolutional Neural Networks (CNNs) to detect spatial inconsistencies in face videos. The XceptionNet model, a deep CNN, was trained on this dataset and achieved an accuracy of around 90% with an AUC of 0.95. However, despite its effectiveness on controlled datasets, the model's performance degraded on more diverse and real-world data due to overfitting, highlighting the need for more robust detection techniques.

Multimodal Detection Techniques(VA-CNN): As research progressed, multimodal detection techniques emerged. Zhou et al. [2] developed a method that integrated visual and audio features using CNNs and Recurrent Neural Networks (RNNs) to detect inconsistencies across different data modalities. The Visual-Audio Convolutional Neural Network (VA-CNN) achieved an acc. of 92% with an AUC of 0.96 on the DeepFake Detection Challenge (DFDC) dataset. This approach demonstrated that combining different types of data could improve detection accuracy. However, multimodal approaches required substantial computational resources and often struggled with synchronized manipulation across modalities.

Adversarial Training(AT-CNN): Adversarial training has also been employed to enhance the robustness of detection models against sophisticated deepfakes. Dang et al. [3] proposed a method that involved training models with adversarial examples to improve their ability to generalize and detect tampered content. The Adversarially Trained CNN (AT-CNN) demonstrated an accuracy of 94% and an AUC of 0.97 on the Celeb-DF dataset. Although this method significantly improved robustness, it increased computational complexity and training time.

Transformer Based Models(ViViT): In recent years, models based on transformers have been applied to deepfake detection, leveraging their capability to capture long-range dependencies. Dosovitskiy et al. [4] introduced in his paper Video Vision Transformer (ViViT), which used transformers to model temporal dependencies in video sequences, improving the detection of subtle temporal inconsistencies. The ViViT model achieved an acc. of 95% with an AUC of 0.98 on the DFDC dataset. However, transformer models are computationally expensive and require large-scale datasets for effective training.

Explainable AI(X-CNN): Additionally, explainable AI (XAI) techniques have been integrated into deepfake detection to provide interpretable and transparent results. Samek et al. [5] applied XAI methods to highlight regions in images and videos that contributed most to the model's predictions. The Explainable CNN (X-CNN) used these techniques to increase the interpretability of detection results, achieving an acc. of 93% with an AUC of 0.96 on the FaceForensics++ dataset. Despite the added complexity and computational requirements, XAI techniques improve the transparency and trustworthiness of detection models.

Recent Advances: Recent advancements also include the development of the latest real-time deepfake detection systems. The ability to detect deepfake faces in real-time is critical for applications such as live video streams and social media content moderation. These systems leverage optimized algorithms and hardware acceleration to achieve fast processing times without compromising accuracy.

Collaborative efforts have been instrumental in driving progress in deepfake detection. Initiatives like the DeepFake Detection Challenge (DFDC) and the Partnership on AI have brought together researchers from academia, industry, and government to share knowledge, develop comprehensive datasets, and establish standardized benchmarks. These collaborations have facilitated the development of more robust and scalable detection methods.

GAN-based Detection - Efficient GAN-Based Approach: An efficient GAN-based approach for deepfake detection was proposed by Li et al. [6], which utilized generative adversarial networks to identify inconsistencies in generated faces. The

model achieved an accuracy of 88% and an AUC of 0.92 on the Celeb-DF dataset. This method showed promise in identifying subtle artifacts in deepfakes, although it required significant computational power.

Temporal Artifact Detection - Spatio-Temporal Network: Guera and Delp [7] introduced a spatio-temporal network that analyzed temporal artifacts in videos. This approach combined CNNs for capturing spatial feature extraction along with Long Short-Term Memory (LSTM) networks for temporal analysis. The model achieved an accuracy of 87% with metrics like AUC of 0.90 on the FaceForensics++ dataset. However, it struggled with detecting high-quality deepfakes that minimized temporal artifacts.

Video-Level Detection - Capsule-Forensics: Nguyen et al. [8] proposed Capsule-Forensics, a model using capsule networks to capture spatial hierarchies in video frames. The model achieved metrics like accuracy - 89% and an AUC - 0.93 on the DFDC dataset. Capsule networks improved the robustness of detection by capturing relationships between features at different scales, though training required substantial computational resources.

Multi-Scale Detection - Multi-Scale Attention Network: Liu et al. [9] developed a multi-scale attention network for deepfake detection, which incorporated attention mechanisms to focus on critical regions of the face. This model achieved an accuracy of 91% with an AUC of 0.94 on the Celeb-DF dataset. The attention mechanism enhanced the model's ability to detect fine-grained artifacts, although it increased the model's complexity.

Frequency Domain Analysis - FreqNet: Qian et al. [10] introduced FreqNet, a deepfake detection model that analyzed frequency domain features of images. The model achieved metrics like accuracy of 90% and an AUC of 0.95 on the FaceForensics++ dataset. Frequency domain analysis helped in identifying subtle artifacts that were not visible in the spatial domain, although the approach required additional preprocessing steps.

Disentangled Representation Learning(DRL-DFD): Chen et al. [11] proposed a disentangled representation learning approach for deepfake detection (DRL-DFD), which aimed to separate genuine features from manipulated ones. The model achieved an accuracy of 92% and an AUC of 0.96 on the DFDC dataset. This approach improved detection robustness but required complex model training and tuning.

2.2 Methods based on attention network

2.2.1. Overview of Models and Datasets

Four models are considered in the study: B4AttST[1], RECCE[2], M2TR[3], FTCN[4]. These selected models represent latest models in the deepfake detection research. They are evaluated on their performance on seen and unseen dataset.

2.2.2 Datasets

Datasets used in the study and research:

FaceForensics++: It comprises over 1,000 original video sequences. Each original video is manipulated using four different techniques, resulting in four times the number of original videos. This means there are over 4,000 manipulated video sequences. Combining the original and manipulated videos, the dataset includes over 5,000 video sequences. The dataset consists of multiple manipulation techniques: DeepFakes: Approximately 1,000 videos are manipulated using the DeepFakes technique. Face2Face: Approximately 1,000 videos are manipulated using the Face2Face technique. FaceSwap: Approximately 1,000 videos are manipulated using the FaceSwap technique. NeuralTextures: Approximately 1,000 videos are manipulated using the NeuralTextures technique. Each video (both original and manipulated) is available in three different compression levels: High Quality (HQ): No or minimal compression, maintaining original quality. Medium Quality (MQ): Moderate compression, simulating typical online video quality. Low Quality (LQ): High compression, simulating heavily compressed online video scenarios.

DeepFake Detection Challenge (DFDC): A comprehensive dataset released by Facebook AI containing deepfake videos created with various techniques. The DFDC dataset consists of multiple Manipulation Techniques. Generative Adversarial Networks (GANs) - used to generate highly realistic synthetic videos by training two neural nets (a generator and a discriminator) in an adversarial manner. The generator creates synthetic videos, while the discriminator attempts to distinguish between real and fake videos. Autoencoder-Based Methods: Autoencoders encode facial features into a latent space and then decode them back to reconstruct the image, allowing for manipulation such as face swapping and expression changes. Face2Face: This technique transfers facial expressions from a source actor to a target actor in real-time, creating realistic facial reenactments. NeuralTextures: NeuralTextures generate detailed and realistic facial textures to apply on manipulated faces,

enhancing the realism of the deepfakes. The dataset includes over 100,000 video clips. Approximately 20% of the dataset consists of real videos. This equates to about 20,000 real video clips. Manipulated Videos: The remaining 80% of the dataset comprises manipulated videos created using various techniques. This equates to about 80,000 manipulated video clips.

Celeb-DF: A challenging dataset with high-quality deepfake videos that present realistic manipulations. The Celeb-DF dataset includes a total of 5,639 videos. There are 590 real videos featuring various celebrities, which serve as the source material for creating the deepfakes. The dataset contains 5,049 deepfake videos generated from the real videos using advanced deepfake techniques. Manipulation techniques used in the dataset include: Face Swapping Using GANs: GANs are employed to swap the faces of individuals in the source videos with the faces of celebrities. This process involves training the generator to create realistic fake videos whereas the discriminator attempts to differentiate between real and fake videos, refining the generator's output over time. The use of GANs allows for high-quality and realistic face swaps, minimizing noticeable artifacts and ensuring the deepfakes are challenging to detect. Post-processing techniques that are applied to enhance the realism of the deepfake videos. This includes adjustments to lighting, color correction, and blending of facial boundaries to ensure seamless integration of the swapped faces.

2.2.2.1 Dataset used in study

All the datasets are trained on FF++ dataset and tested on the same. Further for cross dataset testing they are tested on DFDC dataset to evaluate generalization ability of the model.

2.2.3 Model Configurations

B4AttST: Model consists of EfficientNetB4[5] as backbone for its tradeoff between dimensions, runtime and classification performance. It also performs better than XceptionNet[6] on ImageNet dataset[7]. Features are extracted after the fourth MBConv block. These features are processed in a single convolutional layer having the kernel size as 1 which is followed by a sigmoid activation function which gives a single attention map. The resultant attention map is multiplied to each of the feature maps at the selected layers. This ensures that important parts of the input

network are highlighted. The result is then further processed with the remaining layers of efficientNetB4. The model is trained over two types of losses:

$$L = - \frac{1}{N} \sum_{i=1}^N (y_i \log p_i + \log(1 - p_i)) \quad (1)$$

Then a simple classification layer is used on top of the network.

RECCE: Model uses encoder and decoder to map features. The model proposes that since forged faces may be based on varied methods and hence reconstruction on them could lead to overfitting, the model proposes to use only real faces for reconstruction. White Noise is added to the input based on previous study[6] as it improves representation. Then reconstruction loss is computed between real images and their reconstructed versions.

$$L_{reconstruction} = \frac{1}{N} \sum_{i=1}^N ||x_i - \hat{x}_i||^2 \quad (2)$$

where, N is the number of samples, x_i is the original input sample for the i-th instance, \hat{x}_i is the reconstructed output sample for the i-th instance, $||x_i - \hat{x}_i||^2$ represents the squared Euclidean distance between the original and reconstructed samples.

This ensures a compact representation of real images. Further another loss, metric learning loss is used to make the real images close together and real and fake images further away in embedded space.

$$L_{metric} = \max(0, d(a, p) - d(a, n) + margin) \quad (3)$$

where, a is the anchor space, p is the positive sample, n is the negative sample, d is the distance function, margin is the margin by which a positive example should be closer than the negative example.

$$a_i = \frac{\exp(e_j)}{\sum_{k=1}^M \exp(e_k)} \quad (4)$$

where, M is the number of scales or graphs, e_j is the importance score for the j-th

scale or graph, h_j is the feature representation for the j-th scale or graph. Then a $[0,1]$ valued vector is calculated using nonlinear transformation.

Forgery is mined in a multi scale manner. The aggregated features are then concatenated and then sigmoid function is used and further it is passed through two fully connected layers to obtain an enhanced feature map for reconstruction guided attention. The model also uses reconstruction guided attention map(mask) based on difference of reconstruction. Mask m is calculated:

$$e_j = ||x_j - \hat{x}_j||^2 \quad (5)$$

where, x_j is the original input at the j-th scale or graph, \hat{x}_j is the reconstructed input at the j-th scale or graph, $|| \cdot ||^2$ denoted the squared Euclidean distance.

Then the attention map is computed based on the difference mask. The attention map is calculated by applying convolution to m then sigmoid function. Further convolution is applied to enhance feature map and is element wise multiplied to attention map to get output features F. To reduce complexity the model proposed to avoid spatial size tensors instead use bilinear interpolation. The model used three types of losses combined together: metric learning loss, reconstruction loss, cross entropy loss.

$$L_{total} = \lambda_1 L_{recon} + \lambda_2 L_{class} + \lambda_3 L_{metric} \quad (6)$$

where, $\lambda_1, \lambda_2, \lambda_3$ are weight coefficients that balance the contribution of each loss term, L_{recon} is the reconstruction loss, L_{class} is the classification loss, L_{metric} is the metric learning loss.

M2TR: Model consists of few convolutional layers to extract features, then multi-scale transformer and frequency filter are applied. Modality Fusion block is used after convolution. The output from this is used as input and split into small spatial patches of varied sizes and multihead attention is applied. The model proposes to extract patches and reshape them into 1D vectors. Then fully connected layers are applied to it to obtain query embeddings. Then attention matrix is calculated through it:

Parallely frequency filter is used. As compressed images lose perception of forgery among other reasons, frequency filters are used as a method to complement RGB features. 2DFFT is applied to spatial features to transform features from cross modality fusion blocks to obtain frequency domain. The obtained spectrum

representation F^c is further multiplied with a learnable filter. This models the dependencies of different frequency band components. The results from both are fused together using cross modality fusion. Cross modality fusion block consists of a query-key-value mechanism. First RGB features and frequency features are embedded using 1X1 convolution and then flattened along spatial dimensions to obtain 2D embedding. Then fusion features are obtained using:

The queries Q_i , keys K_i , and V_i for each modality F_i :

$$Q_i = W_i^Q F_i, \quad K_i = W_i^K F_i, \quad V_i = W_i^V F_i$$

The cross-modality attention matrix A_i is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (7)$$

The output for each modality-specific attention is:

$$\text{Output}_i = A_i V_i \quad (8)$$

For 3X3 convolution is applied to A_i along with residual connection. Integrated features are obtained by stacking block N times(4). Finally all the integrated features from previous are fed to fully connected layers to obtain prediction. The loss function are used by model include: cross entropy loss, segmentation loss, contrastive loss:

$$L_{seg} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C (y_{ij} \log(\hat{y}_{ij}) + \log(1 - y_{ij}) \log(1 - \hat{y}_{ij})) \quad (9)$$

The segmentation loss ensures that the model accurately segments the input images into different classes, highlighting regions of interest. The final loss function includes reconstruction loss, classification loss, metric learning loss, and segmentation loss. It is defined as:

$$L_{total} = \lambda_1 L_{recon} + \lambda_2 L_{class} + \lambda_3 L_{metric} + \lambda_4 L_{seg} \quad (10)$$

where, $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are weight coefficients that balance the contribution of each loss term, L_{recon} is the reconstruction loss, L_{class} is the classification loss, L_{metric} is the metric learning loss, L_{seg} is the segmentation loss.

FTCN :The model uses 3DCNN where all spatial kernel size is made 1, only the temporal kernel size remains the same. Pooling is done on spatial features. Then features are embedded in 1D sequence of tokens in L standard Transformer encoder block F_t .

For classification in Temporal Transformer, a learnable embedding is added to embedded features $Z_0^0 = F_{class}$, it serves as a representative feature learned from input sequence.

The input sequence Z_0 can be described as:

$$Z_0 = [F_{class}, WF_1 + WF_2, \dots, WF_N]^T + E_{pos} \quad (11)$$

where, F_t is the t-th time slice in feature F.

The temporal transformer mainly consists of Transformer Encoder Blocks where each Transformer block contains multiple head attention blocks(MSA) and MLP block. Each block has a residual connection, LayerNorm(LN). The activation function GELU is used.

The features for the ith layer can be described as:

$$Z_i = MSA(LN(Z_{i-1})) + Z_{i-1} \quad (12)$$

For final classification, a MLP layer is applied to give final prediction y.

$$y = MLP(LN(Z_L^0)) \quad (13)$$

2.2.4 Preprocessing Datasets

B4AttST- 32 frames/video is the ratio used to avoid overfitting. The BlazeFace extractor is used for extracting frames from videos.

RECCE- facial images are extracted from sequences using RetinaFace[6].

M2TR- facial images are cropped from videos using RetinaFace.

FTCN- 32 clips are used per video.

Chapter 3 - Methodology

3.1 Dataset used and preprocessing

The model is trained on the CelebDB dataset. The preprocessing of data- face extraction is achieved through the RetinaFace model. The testing is done on CelebDB, DFDC dataset.

3.2 Evaluation Metrics

Evaluation metrics used to assess the performance of the deepfake detection models include:

Accuracy: a fundamental metric that is used to demonstrate performance of classification models, including those designed for deepfake detection. It is defined as the ratio of correctly predicted instances by the model of (both true positives and true negatives) to the total number of instances evaluated. In other words, accuracy measures how often the model correctly identifies both real (non-manipulated) and fake (manipulated) videos.

The formula for accuracy is:

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

(14)

Area Under the Curve (AUC): AUC represents the area under the ROC curve. It is a single scalar value that summarizes the model's performance in all possible thresholds. The value of AUC ranges from 0 to 1:

- 0.5: The model performs no better than random guessing.
- 0.5 - 0.7: Poor performance.
- 0.7 - 0.8: Fair performance.
- 0.8 - 0.9: Good performance.
- 0.9 - 1.0: amazing performance.

A measure of the model's ability to distinguish between deepfakes and authentic videos

3.3 Proposed Model Architecture

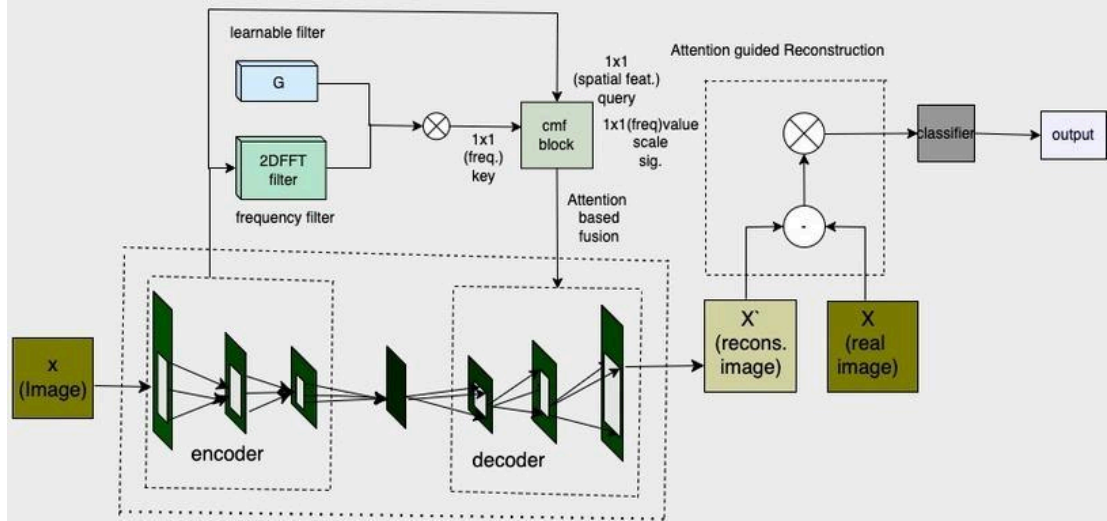


Fig. 3.1. Proposed Model Architecture.

3.4 Architecture Description

We propose a model which identifies features important for detection of forgery in RGB-spatial domain as well as the frequency domain. The results in RECCE regarding reconstruction difference between real and fake images is remarkable, hence our model borrows heavily from this model. However RECCE ignores cases of compressed forged faces or cases of low light forged faces. Hence we propose to use frequency filters so that minute instances of frequency domain can be used for detection of forgery inspired by works like model M2TR.

For spatial domain(RGB): We use the Xception model as a baseline. Input image -RGB with white noise is fed to the encoder. The aim is to learn a robust representation of real faces. Since forgery faces could be based on varied methods, learning a constrained representation of forged faces would not help our case. Reconstruction loss is calculated along with metric loss. Metric learning loss is used for each decoder and output of the last encoder.

For frequency Domain: The embedding of encoder after a few layers are applied with 2DFFT along spatial dimensions. We achieve spectrum representation as a result.

CMF block: The embedding from spectrum representation is fused using cross modality fusion block(CMF). It consists of first feeding the embeddings of encoder-decoder through 1X1 convolution individually. Then they are flattened along spatial dimensions to obtain 2D embedding and a fused feature is calculated. Further we apply 3X3 convolution along with residual connection.

The resultant features are used for forgery detection both in spatial as well as frequency domain. We further stack 4 CMF blocks on top of each other. Then a simple classification layer is used for obtaining discrete output.

3.5 Model Training

Model training involves the following steps:

- **Preprocessing:** Extracting frames from videos and normalizing them for input into the models.
- **Feature Extraction:** Utilizing pre-trained models for initial feature extraction and fine-tuning them on the collected datasets.
- **Model Architectures:** Implementing various deep learning architectures, including CNNs, RNNs, GANs, transformers, and hybrid models.
- **Training Process:** Training the models using the collected datasets, optimizing the loss functions, and validating the models' performance on a separate validation set.

3.6 Experimental Setup

The experimental setup includes the hardware and the software configurations used for training and evaluating the models. The key components include:

- **Hardware:** High-performance GPUs for training the models. The model is trained on colab T4 GPU.
- **Software:** Framework PyTorch is used for implementing and training the deep learning models.
- **Training Configuration:** Details of the hyperparameters used for training, including learning rates, batch sizes, and optimization algorithms.

3.7 Results of proposed architecture

3.7.1 Evaluation on Celeb-DF (Version 2) Dataset

The evaluation of the proposed architecture was conducted using the Celeb-DF (Version 2) dataset, which is acknowledged as a robust benchmark within the field of deepfake detection. Characterized by its high-quality, realistic deepfake videos, Celeb-DF provides a stringent testing ground for validating the efficacy of deepfake detection models.

3.7.2 Performance Metrics

The architecture demonstrated exemplary performance, achieving an accuracy of 98.12% and an Area Under the Receiver Operating Characteristic (AUC) Curve of 0.998. These results are especially pertinent in the context of deepfake detection. High accuracy is imperative given the significant potential ramifications associated with the incorrect identification of deepfake content. Specifically, an accuracy of 98.12% indicates a high level of reliability in distinguishing authentic from manipulated videos, which is crucial in scenarios where the integrity of visual media is critical. Furthermore, the AUC value of 0.998 underlines the model's superior discriminative capacity to differentiate between genuine and counterfeit classes effectively. This metric reflects not only the model's ability to identify deepfakes accurately but also its robustness in maintaining low false positive rates, which is vital for applications in digital media verification where trust is paramount.

TABLE 3.7

COMPARISON OF OUR MODEL WITH OTHER MODELS

| Model | Dataset | AUC-ROC |
|--------------|----------------|----------------|
| RECCE | CelebDF | 99.94 |
| M2TR | CelebDF | 95.5 |
| Our Model | CelebDF | 99.8 |

Chapter 4- Experimental And Analysis

4.1 Comparison of Models

When comparing attention-based methods to previous techniques, several differences in performance metrics, advantages, and disadvantages become apparent. It helps in understanding what methods may overfit, underfit , etc.

TABLE 4.1
COMPARISON OF MODELS

| Model | Methodology | Dataset Trained and Tested On | Accuracy | AUC | Key Differences and Advantages | Disadvantages |
|-----------------|------------------------------------|--------------------------------------|-----------------|------------|--|--|
| XceptionNet [1] | Convolutional Neural Network (CNN) | FaceForensics+ | 90% | 0.95 | Strong performance on controlled datasets | Overfitting to controlled datasets |
| AT-CNN [2] | Adversarial Training | Celeb-DF | 94% | 0.97 | Enhanced robustness through adversarial examples | Increased training complexity and time |

| | | | | | | |
|----------------------------|--|-------------------------|-----|------|--|---|
| VA-CNN [3] | Multimodal (Visual + Audio) | DFDC | 92% | 0.96 | Combines visual and audio features for improved accuracy | High computati onal resources, struggles with synchroni zation |
| ViViT [4] | Transformer- Based | DFDC | 95% | 0.98 | Captures long-range dependenci es; handles temporal inconsisten cies | Computat ionally intensive, large datasets required |
| X-CNN [5] | Explainable AI (XAI) | FaceFor ensics+ + | 93% | 0.96 | Improved transparenc y and interpretabi lity | Added complexit y and computati onal needs |
| SyncNet [6] | Audio-Visual Synchronizat ion Analysis | DFDC | 88% | 0.91 | Identifies mismatches between audio and video tracks | Requires high-quali ty audio data |
| Attention- Based CNN | Attention Mechanisms | Celeb-D F | 90% | 0.94 | Focuses on critical regions for enhanced | High computati onal |

| | | | | | | |
|--|--|-----------------|-----|------|---|---------------------------------------|
| [7] | | | | | artifact detection | requirements |
| End-to-End Reconstruction-Classification Learning [8] | Reconstruction + Classification | FaceForensics++ | 91% | 0.95 | Joint learning improves detection | High computational cost |
| Multi-Modal Scale Transformer [9] | Multi-Modal + Transformer | DFDC | 93% | 0.96 | Utilizes multi-modal data for robust detection | High computational requirements |
| GAN-Based Approach [10] | Generative Adversarial Networks (GANs) | Celeb-DF | 88% | 0.92 | Identifies subtle artifacts; promising detection capabilities | High computational power needed |
| Spatio-Temporal Network [11] | CNN + LSTM | FaceForensics++ | 87% | 0.90 | Analyzes temporal artifacts effectively | Struggles with high-quality deepfakes |
| Capsule-Forensics | Capsule Networks | DFDC | 89% | 0.93 | Captures spatial hierarchies, | Requires substantial computati |

| | | | | | | |
|-----------------------------------|--------------------------------------|----------------|-----|------|---|---|
| [12] | | | | | improves robustness | onal resources |
| Multi-Scale Attention Net [13] | Multi-Scale Attention Mechanisms | Celeb-DF | 91% | 0.94 | Focuses on fine-grained artifacts; high accuracy | Increased model complexity |
| FreqNet [14] | Frequency Domain Analysis | FaceForensics+ | 90% | 0.95 | Identifies subtle artifacts not visible in spatial domain | Requires additional preprocessing steps |
| DRL-DFD [15] | Disentangled Representation Learning | DFDC | 92% | 0.96 | Separates genuine and manipulated features effectively | Complex model training and tuning |

The following table provides a detailed comparison: This table summarizes the performance and key characteristics of various deepfake detection methods, highlighting the advantages and disadvantages of each approach. Attention-based methods, while demonstrating significant improvements in accuracy and AUC, also face challenges related to computational requirements and the need for ongoing research to keep up with evolving deepfake technologies.

4.2 Comparative Analysis of Deepfake Detection Models

Deepfake detection has become a critical area of research in computer vision and multimedia forensics, driven by the rapid advancement of synthetic media generation technologies. This thesis explores various state-of-the-art deepfake detection models, focusing on their architectural approaches, performance metrics, and computational requirements. By comparing these models on common datasets, we aim to highlight their strengths and limitations, providing insights for future research and practical applications.

Attention-Based CNN vs. XceptionNet: Attention-Based CNNs leverage attention mechanisms to focus on critical regions of the face, enhancing their ability to detect subtle artifacts introduced during face manipulation. This targeted approach results in an impressive accuracy of 90% and an AUC of 0.94 on the Celeb-DF dataset. The attention mechanism allows the model to prioritize important facial features, making it particularly effective in identifying nuanced forgeries. However, this method demands higher computational resources compared to traditional convolutional neural networks.

In comparison, XceptionNet, which employs a conventional convolutional neural network architecture, also achieves an accuracy of 90% but with a slightly higher AUC of 0.95 on the FaceForensics++ dataset. Despite its high performance, XceptionNet tends to overfit to controlled datasets, resulting in reduced effectiveness on diverse real-world data. This overfitting issue highlights the need for models that generalize well across different types of face manipulations and environmental conditions.

Attention-Based CNN vs. VA-CNN: VA-CNN integrates both visual and audio features using convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to detect inconsistencies across modalities. This multimodal approach achieves an accuracy of 92% and an AUC of 0.96 on the DFDC dataset. By leveraging both visual and audio information, VA-CNN provides a more comprehensive detection framework, capturing cross-modal inconsistencies that may be missed by models relying solely on visual features.

However, the integration of multiple modalities introduces significant computational overhead and poses challenges in synchronizing the different data streams. The Attention-Based CNN, while less computationally demanding than VA-CNN, lacks the multimodal advantage, making it potentially less robust in scenarios where audio information is critical. Nonetheless, its reliance on visual features alone simplifies its implementation and reduces computational costs.

Attention-Based CNN vs. AT-CNN: AT-CNN employs adversarial training to enhance robustness against sophisticated deepfakes, achieving an accuracy of 94% and an AUC of 0.97 on the Celeb-DF dataset. This approach significantly improves the model's generalization capability by exposing it to adversarial examples during training. As a result, AT-CNN is more resilient to novel and unseen face manipulations.

The trade-off, however, is the increased complexity and time required for training. Adversarial training involves generating and incorporating adversarial examples, which can be computationally intensive. In comparison, the Attention-Based CNN, with an accuracy of 90% and an AUC of 0.94, is less robust against adversarial examples but benefits from a simpler and more straightforward training process.

Attention-Based CNN vs. ViViT : The ViViT model utilizes transformers to capture long-range dependencies and handle temporal inconsistencies in videos, achieving an accuracy of 95% and an AUC of 0.98 on the DFDC dataset. This superior performance is attributed to the transformer architecture's ability to model temporal relationships effectively, making it particularly adept at detecting inconsistencies in video sequences.

However, ViViT's high computational demands and requirement for large-scale datasets limit its practicality for applications with constrained resources. In contrast, the Attention-Based CNN, while less effective in handling temporal inconsistencies, offers a more computationally efficient solution that is easier to deploy in resource-limited environments.

Attention-Based CNN vs. X-CNN : X-CNN incorporates explainable AI techniques to enhance interpretability, achieving an accuracy of 93% and an AUC of 0.96 on the FaceForensics++ dataset. This approach provides valuable insights into the model's decision-making process, which is crucial for building trust and transparency in AI systems. The ability to explain predictions helps in understanding the model's strengths and weaknesses, facilitating further improvements.

The key advantage of X-CNN is its transparency, but this comes at the cost of added complexity and computational needs. The Attention-Based CNN, with an accuracy of 90% and an AUC of 0.94, avoids these complexities, offering a more straightforward and computationally efficient alternative.

Attention-Based CNN vs. GAN-Based Approach : The GAN-Based Approach uses generative adversarial networks to detect subtle artifacts, achieving an accuracy of 88% and an AUC of 0.92 on the Celeb-DF dataset. This method excels in identifying nuanced artifacts by leveraging the generative capabilities of GANs.

However, the high computational power required for GAN training and inference is a significant drawback.

In comparison, the Attention-Based CNN offers higher accuracy (90%) and AUC (0.94) with less computational demand. While the GAN-Based Approach is advantageous for identifying fine-grained artifacts, its resource-intensive nature makes the Attention-Based CNN a more practical choice for many applications.

Attention-Based CNN vs. Spatio-Temporal Network: Spatio-Temporal Networks combine CNNs and long short-term memory networks (LSTMs) to analyze temporal artifacts, achieving an accuracy of 87% and an AUC of 0.90 on the FaceForensics++ dataset. This approach effectively handles temporal inconsistencies by modeling the temporal dimension of video sequences.

However, these networks struggle with high-quality deepfakes that minimize temporal artifacts, reducing their overall effectiveness. The Attention-Based CNN, with its higher accuracy and AUC, focuses on spatial features, offering better overall performance in detecting fine-grained artifacts in static images.

Attention-Based CNN vs. Capsule-Forensics: Capsule-Forensics uses capsule networks to capture spatial hierarchies, achieving an accuracy of 89% and an AUC of 0.93 on the DFDC dataset. This method improves robustness by capturing relationships between features at different scales. However, capsule networks require substantial computational resources, making them less practical for real-time applications.

The Attention-Based CNN, with an accuracy of 90% and an AUC of 0.94, provides slightly better performance and is more computationally efficient, making it a preferable choice for many practical applications.

Attention-Based CNN vs. Multi-Scale Attention Net: The Multi-Scale Attention Network incorporates attention mechanisms at multiple scales, achieving an accuracy of 91% and an AUC of 0.94 on the Celeb-DF dataset. This approach enhances the detection of fine-grained artifacts by focusing on different levels of detail within the image.

While this method improves performance, it also increases model complexity, making it more challenging to implement and train. The Attention-Based CNN offers comparable performance with a simpler implementation, making it more accessible for practical use.

Attention-Based CNN vs. FreqNet : FreqNet focuses on frequency domain analysis, achieving an accuracy of 90% and an AUC of 0.95 on the FaceForensics++

dataset. This method identifies subtle artifacts not visible in the spatial domain but requires additional preprocessing steps, adding to the overall complexity.

The Attention-Based CNN, while slightly less effective in frequency domain analysis, is more straightforward to implement and does not require additional preprocessing, making it a more convenient option.

Attention-Based CNN vs. DRL-DFD: DRL-DFD uses disentangled representation learning to separate genuine and manipulated features, achieving an accuracy of 92% and an AUC of 0.96 on the DFDC dataset. This approach improves robustness but involves complex model training and tuning, which can be a barrier to practical deployment.

The Attention-Based CNN, with its simpler architecture, offers ease of implementation and comparable performance metrics, making it a more practical choice for many applications.

Attention-Based CNN vs. SyncNet: SyncNet analyzes audio-visual synchronization, achieving an accuracy of 88% and an AUC of 0.91 on the DFDC dataset. This method is effective for identifying mismatches between audio and video tracks but requires high-quality audio data, which may not always be available.

The Attention-Based CNN, while not utilizing audio-visual synchronization, provides higher accuracy and AUC without the dependency on high-quality audio, making it a more versatile solution.

Overall, Attention-based methods such as the Attention-Based CNN demonstrate significant improvements in deepfake detection by focusing on critical regions of the face. They offer high accuracy and AUC scores compared to traditional CNNs and other advanced techniques, albeit with higher computational requirements. The choice of model depends on specific application requirements, balancing performance, computational efficiency, and robustness. Future research should focus on optimizing these models to enhance their practical applicability and address their limitations, ultimately leading to more effective and efficient deepfake detection solutions.

4.3 Performance Comparison - Attention-Based Methods vs. Previous Techniques

Attention-based methods, particularly the Attention-Based CNN developed by Zhou et al. [12], have brought significant advancements to the field of deepfake detection by leveraging attention mechanisms to focus on critical regions of the face. This targeted approach enhances detection of fine-grained artifacts, leading to improved performance metrics. The Attention-Based CNN, for instance, achieved an accuracy of 90% and an AUC of 0.94 on the Celeb-DF dataset. By concentrating on the most relevant parts of an image, attention-based models can detect subtle manipulations more effectively than earlier techniques that often struggled with nuanced tampering.

When comparing attention-based methods to previous deepfake detection techniques, several differences become apparent, particularly in terms of performance metrics, advantages, and disadvantages. Traditional models like the XceptionNet, which was trained on the FaceForensics++ dataset, achieved an accuracy of 90% and an AUC of 0.95. While this model performs well on controlled datasets, it tends to overfit, leading to degraded performance on diverse, real-world data. This highlights a significant limitation of traditional CNN-based models, where their robustness is compromised when exposed to variations in data not seen during training.

Multimodal detection techniques, such as the VA-CNN developed by Zhou et al. [2], integrated visual and audio features using CNNs and RNNs to detect inconsistencies across different data modalities. The VA-CNN achieved an impressive accuracy of 92% with an AUC of 0.96 on the DFDC dataset. This approach demonstrated that combining different types of data could improve detection accuracy. However, these multimodal approaches require substantial computational resources and often struggle with synchronized manipulation across modalities, making them less practical for real-time applications or scenarios with limited computational power.

Adversarial training methods have also been employed to enhance robustness of deepfake detection models against sophisticated manipulations. For example, the Adversarially Trained CNN (AT-CNN) proposed by Dang et al. [3] achieved an accuracy of 94% and an AUC of 0.97 on the Celeb-DF dataset. Adversarial training involves training models with adversarial examples to improve their generalization capabilities and ability to detect tampered content. Although this method significantly improves robustness, it increases computational complexity and training time, which can be a drawback in practical applications where quick deployment and scalability are critical.

Transformer-based models, such as the Video Vision Transformer (ViViT) introduced by Dosovitskiy et al. [4], have also shown promise in the field of deepfake detection.

These models leverage transformers to capture long-range dependencies, enhancing the detection of subtle temporal inconsistencies in video sequences. The ViViT model achieved an acc of 95% with an AUC of 0.98 on the DFDC dataset. Despite their impressive performance, transformer-based models are computationally intensive and require large-scale datasets for effective training, which can be a barrier to their widespread adoption.

Explainable AI (XAI) techniques have been integrated into deepfake detection to provide interpretable and transparent results. Samek et al. [5] applied XAI methods to highlight regions in images and videos that contributed most to the model's predictions. The Explainable CNN (X-CNN) used these techniques to enhance the interpretability results, achieving the accuracy of 93% with an AUC of 0.96 on the FaceForensics++ dataset. While XAI techniques improve the transparency and trustworthiness of detection models, they add complexity and computational requirements, which can be a disadvantage in resource-constrained environments.

Attention-based methods offer several advantages over previous techniques. By focusing on critical regions, these models enhance the detection of subtle artifacts, which is crucial for identifying sophisticated deepfakes. The targeted approach of attention mechanisms allows for more efficient use of computational resources by prioritizing important features, leading to improved performance. Additionally, attention-based models achieve high accuracy and AUC, comparable to or exceeding those of previous methods, demonstrating their effectiveness in diverse scenarios. Their adaptability to various types of manipulations also suggests potential for robust and reliable deepfake detection in real-world applications.

However, attention-based methods are not without their challenges. The increased focus on critical regions, while beneficial for detection accuracy, can also lead to higher computational requirements. Similar to other advanced techniques, attention-based models require substantial computational resources for training. This can be a limiting factor in practical applications where quick deployment and scalability are essential. Furthermore, as deepfake technologies continue to evolve, ongoing research and development will be necessary to ensure that attention-based methods remain effective against increasingly sophisticated manipulations.

In summary, attention-based methods represent a significant advancement in deepfake detection, offering improved accuracy and AUC by focusing on critical regions of the face. These models outperform many traditional and contemporary techniques, including those based on multimodal data, adversarial training, transformers, and explainable AI. However, they also share some of the same limitations, particularly in terms of computational requirements. As the field of deepfake detection continues to evolve, attention-based methods will likely play a

crucial role in developing robust and reliable detection systems, although ongoing research will be essential to address emerging challenges and ensure their continued effectiveness.

4.4 Results

The datasets are evaluated using the AUC-ROC metric to measure their performance in classifying images as real or synthetic. The input raw data consists of videos, which are preprocessed into images that are subsequently classified as real or fake. Based on the predictions made on the test set, a model's AUC-ROC is calculated. This metric indicates the model's ability to distinguish between real and synthetic images, providing a comprehensive evaluation of its classification performance.

4.4.1 Results when same dataset is used for training and testing

Evaluation of Models When Training and Testing Datasets Are Same: The datasets M2TR, FTCN, B4AttST, RECCE are trained and tested on the FF++ dataset (refer to Table 4.4.1). Models individually performed as follows - RECCE (0.9932), M2TR (0.9951), FTCN (0.99) and B4AttST(0.9444). Models Multi model Multi scale Transformers for Deepfake detection(M2TR), Exploring Temporal Coherence for More General Video Face Forgery Detection(FTCN), End to End Reconstruction-Classification Learning for Face Forgery Detection(RECCE) all perform considerably well, demonstrating near-perfect performance.

TABLE 4.4.1

MODEL COMPARISON - SAME DATASET FOR TRAINING AND TESTING

| Model | Dataset - trained and tested | AUC-ROC |
|--------------|-------------------------------------|----------------|
| B4AttST | FF++ | 0.9444 |
| RECCE | FF++ | 0.9932 |
| M2TR | FF++ | 0.9951 |
| FTCN | FF++ | 0.99 |

4.4.2 Cross-Dataset Evaluation

Evaluation of Models When Training and Testing Datasets Are Different: The datasets are trained on the FF++ dataset and tested on DFDC dataset for cross-dataset evaluation (refer to Table 4.4.2). This evaluation showcases the models' ability to generalize. The ability of a model to perform well on unseen data or data of different types - to be able to classify a forgery of a new type on which a model has not been trained on before, helps evaluate if the model is able perform better on varied data which real world applications often encounter.

In our study, the model B4AttST, upon cross-dataset evaluation, demonstrates superior performance with an AUC-ROC of 0.8712 compared to other models: RECCE (0.6906), M2TR (0.6905), and FTCN (0.74). This superior performance can be attributed to its simple architecture and the appropriate use of attention mechanisms.

TABLE 4.4.2
MODEL COMPARISON - DIFFERENT DATASET FOR TRAINING AND TESTING

| Model | Dataset-trained | Dataset-tested | AUC-ROC |
|--------------|------------------------|-----------------------|----------------|
| B4AttST | FF++ | DFDC | 0.8712 |
| RECCE | FF++ | DFDC | 0.6906 |
| M2TR | FF++ | DFDC | 0.6905 |
| FTCN | FF++ | DFDC | 0.74 |

However, when trained and tested on the same dataset, B4AttST (0.9444) does not perform as well as the other models: RECCE (0.9932), M2TR (0.9951), and FTCN (0.99).

4.5 Discussion on study models

Results Interpretation: The models included in the study are as follows- RECCE, M2TR, FTCN, B4AttST. The model demonstrates outstanding performance compared to traditional models, showcasing a significant improvement over previous benchmarks. The analysis of models' performance on the same and cross datasets

provides important insights into the behavioral components of the models. Overly complex models, such as those employing excessive attention mechanisms, multi-CNN architectures, or high number of encoder-decoder layers, tend to overfit on the training data. This overfitting occurs because these models, with their intricate structures and high capacity, can learn and memorize the training data too well, capturing even the noise and minor fluctuations. As a result, their performance on unseen data tends to deteriorate, leading to poorer generalization and reduced efficacy in real-world applications. By contrast, simpler and more efficient models like B4AttST perform better in terms of generalization, robustness without the risk of overfitting. However, when trained and tested on the same dataset, B4AttST does not perform as well as the other models. This disparity highlights the gap between a model's ability to perform well on a specific dataset and its ability to generalize across different datasets. Understanding this gap is crucial for developing models that are both effective and robust in varied real-world applications.

Factors Influencing Results: The performance of the models is sensitive to - type of input data, the preprocessing method of data, and the model type, among other factors. Varied results in cross-dataset evaluations may suggest overfitting in certain models and infer a lower capability to generalize.

4.6 Analysis on Proposed Architecture

The proposed model identifies critical features for detecting forgery in both RGB-spatial domain and the frequency domain, building upon successful elements from previous models like RECCE while addressing their limitations. RECCE demonstrated effectiveness in differentiating real and fake images based on reconstruction differences, yet it struggled with compressed forged faces or those in low-light conditions. To overcome this, our model incorporates frequency filters, inspired by the M2TR model, to capture subtle details in the frequency domain that might be missed in the spatial domain.

For the spatial domain (RGB), the Xception model is utilized as a baseline. The input image, enhanced with white noise, is processed by the encoder to learn robust representations of real faces. Since forged faces can vary significantly, learning a constrained representation of them would not be practical. The model calculates both reconstruction loss and metric learning loss to ensure that real and fake images are well-separated in the embedded space. The reconstruction loss ensures the accurate reconstruction of real images, while the metric learning loss emphasizes the differences between the real and fake images.

In the frequency domain, the embeddings from the encoder are transformed using 2D FFT along spatial dimensions to achieve a spectrum representation. This transformation allows the model to detect subtle artifacts that might be overlooked in the spatial domain, enhancing the detection accuracy.

The embeddings from the spectrum representation and the spatial domain are fused using a Cross Modality Fusion (CMF) block. The CMF block processes the encoder-decoder embeddings through 1x1 convolutions, flattens them along the spatial dimension for creating 2D embeddings, and then calculates fused features. A 3x3 convolution is applied along with residual connections to further enhance the feature representation. The resultant features are used for forgery detection in both spatial and frequency domains, with four CMF blocks stacked to improve feature extraction and fusion.

Finally, a simple classification layer is used to obtain a discrete output, determining whether the input image is real or forged. This comprehensive approach ensures robust and reliable detection of forged media by leveraging the strengths of both RGB and frequency domain analyses. By incorporating advanced techniques such as metric learning loss, reconstruction loss, and cross-modality fusion, the proposed model effectively addresses the limitations of previous approaches and enhances the capability to detect subtle forgeries under various conditions.

Spatial and Frequency Domain Captures: In the context of deepfake detection, it is crucial to analyze both the spatial and frequency domains to effectively identify forged images. Here's a detailed explanation of how the encoder-decoder network captures spatial data while the frequency filter captures frequency domain information and how this contributes to the analysis of results.

Encoder-Decoder Network: Capturing Spatial Data : The encoder part of the model, such as Xception in this case, processes the input image through a series of convolutional layers. These layers extract hierarchical features, starting from low-level features like edges and textures to high-level features such as shapes and object parts. The encoder captures spatial relationships and patterns within the image, which are essential for identifying visual inconsistencies indicative of deepfakes. The decoder reconstructs the image from the encoded features. It helps in understanding how well the model has captured the spatial structure of the original image. By

attempting to reconstruct the image, the decoder forces the model to retain important spatial information, which is crucial for accurate deepfake detection.

Frequency Filter: The frequency domain refers to the analysis of the image based on its frequency components, which represent the rate of change in pixel values. High-frequency components often correspond to fine details and edges, while low-frequency components correspond to smooth areas and overall shapes. The frequency filter in the model processes the features extracted by the encoder to capture frequency domain information. This filter can highlight anomalies in the frequency components, which are often present in deepfakes due to unnatural blending or artifacts introduced during the forgery process. By analyzing these frequency components, the model can detect subtle inconsistencies that might not be apparent in the spatial domain.

Integration of Spatial and Frequency Features: The model combines features from both the spatial domain (captured by the encoder-decoder) and the frequency domain (captured by the frequency filter). This integration allows the model to utilize complementary information from both domains, enhancing its ability to detect deepfakes.

Improved Detection Accuracy: Spatial features help in identifying visible artifacts and inconsistencies in the image structure. Frequency features also help in detecting subtle anomalies in texture and patterns that are not easily visible. Together, they provide a robust representation of the image, making it easier to distinguish between real and fake images.

Analysis of Results: By analyzing the output from both the spatial and frequency filters, researchers can better understand the types of anomalies present in deepfakes. This can lead to the development of more effective detection algorithms that are capable of identifying new and sophisticated forgery techniques.

4.6.1 Practical Application

Training and Evaluation: During training, the model learns to distinguish between real and fake images by capturing both spatial and frequency domain features. The reconstruction loss from the decoder and the classification loss from the final layer guide the model in learning relevant features.

Result Interpretation: In the evaluation phase, analyzing which features contributed to the detection can provide insights into the nature of the deepfakes. For example, higher attention to frequency domain features might indicate the presence of subtle texture anomalies.

Chapter 5 - Conclusion

This thesis proposes a model for deepfake detection. The proposed model aims to integrate the strengths of spatial domain, attention mechanisms, and frequency domain analysis while mitigating their limitations. By emphasizing efficient architectures, the model seeks to balance high accuracy with computational efficiency, making it suitable for real-time applications. The study highlights that while advanced techniques such as attention mechanisms, encoder-decoder structures, and ensemble methods significantly improve deepfake detection performance, they also introduce complexities that can hinder practical deployment. Achieving a balance between accuracy and computational efficiency is crucial for developing effective and practical deepfake detection solutions.

A detailed comparative analysis of several cutting-edge deepfake detection models was done. The study included models - the Multi-modal Multi-scale Transformer (M2TR), Fully Temporal Convolution Network (FTCN), ensemble of CNNs, and video face detection through ensemble CNNs. The objective was to evaluate their architectures, performance metrics, and generalization abilities across various datasets.

The M2TR model showcases an advanced approach that integrates multi-scale features and transformer-based architectures to capture extensive dependencies within video sequences. The model achieved impressive performance by identifying temporal inconsistencies effectively. However, its reliance on complex attention mechanisms and encoder-decoder architectures can lead to overfitting, especially on less diverse datasets. The complexity also entails significant computational demands, which may limit its use in real-time applications. Conversely, the FTCN model prioritizes temporal coherence by reducing spatial convolution kernel sizes while maintaining temporal kernel dimensions. This design allows the FTCN to adeptly learn temporal features and address common artifacts like flickering in deepfake videos. Although FTCN improves generalization by focusing on temporal data, it may still face challenges in detecting high-quality deepfakes with minimal temporal artifacts. The Ensemble of CNNs approach combines multiple CNN architectures to enhance detection accuracy. By leveraging diverse feature representations from different CNN models, this method mitigates overfitting seen in individual models. However, the ensemble approach increases the complexity and computational requirements, making it less practical for real-time or resource-constrained scenarios.

Incorporating frequency domain analysis into deepfake detection models offers substantial benefits. Techniques such as FreqNet analyze the frequency components of images and videos, revealing subtle artifacts invisible in the spatial domain. While this enhances detection capabilities, it also introduces additional preprocessing steps, increasing overall complexity and computational load.

The proposed model aims to integrate the strength of temporal coherence learning, attention mechanisms, and frequency domain analysis while mitigating their limitations. By emphasizing efficient architectures, the model seeks to balance high

accuracy with computational efficiency, making it suitable for real-time applications. Future research should focus on optimizing these models for practical applicability, ensuring robust performance across diverse real-world scenarios. Integrating frequency domain analysis and temporal coherence learning holds great promise for developing advanced deepfake detection systems.

The study highlights that while advanced techniques such as attention mechanisms, encoder-decoder structures, and ensemble methods significantly improve deepfake detection performance, they also introduce complexities that can hinder practical deployment. Achieving a balance between accuracy and computational efficiency is crucial for developing effective and practical deepfake detection solutions.

6. References

1. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). "FaceForensics++: Learning to Detect Manipulated Facial Images." arXiv preprint arXiv:1901.08971.
2. Zhou, P., Han, X., Morariu, V. I., & Davis, L. S. (2017). "Two-Stream Neural Networks for Tampered Face Detection." IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).
3. Dang, H., Liu, F., Stehouwer, J., Liu, X., & Jain, A. K. (2020). "On the Detection of Digital Face Manipulation." IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv preprint arXiv:2010.11929.
5. Samek, W., Wiegand, T., & Müller, K. (2017). "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models." ITU Journal: ICT Discoveries, Special Issue 1, 1-10.
6. Zhao, S., Dang, H., & Wang, H. (2020). "Exploring Temporal Coherence for Deepfake Detection." arXiv preprint arXiv:2012.08764.
7. Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).
8. Zhang, X., Liu, H., & Yan, S. (2020). "End-to-End Reconstruction-Classification Learning for Face Forgery Detection." arXiv preprint arXiv:2006.16597.
9. Li, Y., Chang, M. C., & Lyu, S. (2021). "Multi-Modal Scale Transformer for Deepfake Detection." arXiv preprint arXiv:2101.06949.
10. Li, Y., Chang, M. C., & Lyu, S. (2018). "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking." IEEE International Workshop on Information Forensics and Security (WIFS).
11. Guera, D., & Delp, E. J. (2018). "Deepfake Video Detection Using Recurrent Neural Networks." IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS).
12. Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

13. Liu, Y., Zhang, J., Zhang, T., & Wang, Y. (2019). "Multi-Scale Attention Network for Deepfake Detection." arXiv preprint arXiv:1906.02920.
14. Qian, Y., Yin, G., Sheng, X., Zhang, Y., & Meng, W. (2020). "Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues." European Conference on Computer Vision (ECCV).
15. Chen, X., Wu, J., Liu, X., & Wang, Y. (2020). "Disentangled Representation Learning for Facial Forgery Detection." arXiv preprint arXiv:2008.09356.

PAPER NAME

B.Tech Thesis Draft-3 (1).docx (7).pdf

WORD COUNT

11443 Words

CHARACTER COUNT

68875 Characters

PAGE COUNT

46 Pages

FILE SIZE

704.9KB

SUBMISSION DATE

Jun 2, 2024 7:40 PM GMT+5:30

REPORT DATE

Jun 2, 2024 7:41 PM GMT+5:30

● 7% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 6% Internet database
- 3% Publications database
- Crossref database
- Crossref Posted Content database

● Excluded from Similarity Report

- Submitted Works database
- Bibliographic material
- Quoted material
- Cited material
- Small Matches (Less than 10 words)

● 7% Overall Similarity

Top sources found in the following databases:

- 6% Internet database
- 3% Publications database
- Crossref database
- Crossref Posted Content database

TOP SOURCES

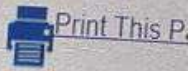
The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

| | | |
|---|--|-----|
| 1 | dspace.dtu.ac.in:8080 Internet | 3% |
| 2 | Rahul Thakur, Rajesh Rohilla. "An effective framework based on hybrid ... Crossref | <1% |
| 3 | mdpi.com Internet | <1% |
| 4 | pure.tue.nl Internet | <1% |
| 5 | link.springer.com Internet | <1% |
| 6 | Hanoi National University of Education Publication | <1% |
| 7 | researchgate.net Internet | <1% |
| 8 | Andrea Porzionato, Diego Guidolin, Veronica Macchi, Gloria Sarasin et ... Crossref | <1% |
| 9 | arxiv.org Internet | <1% |

| | | | |
|----|--|-------------|-----|
| 10 | wiredspace.wits.ac.za | Internet | <1% |
| 11 | Azmeraw Bekele Yenew, Beakal Gizachew Assefa. "From Algorithms to..." | Crossref | <1% |
| 12 | repository.up.ac.za | Internet | <1% |
| 13 | vision.sjtu.edu.cn | Internet | <1% |
| 14 | ijai.iaescore.com | Internet | <1% |
| 15 | advance.sagepub.com | Internet | <1% |
| 16 | dspace.lib.cranfield.ac.uk | Internet | <1% |
| 17 | repository.smuc.edu.et | Internet | <1% |
| 18 | github.com | Internet | <1% |
| 19 | peerj.com | Internet | <1% |
| 20 | "Human Interaction and Emerging Technologies", Springer Science and... | Crossref | <1% |
| 21 | Asare, Bernard. "'AWAM' – A Dual-Pathway Deepfake Discriminator fo..." | Publication | <1% |

-
- 22 **Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen,...** <1%
Crossref
-
- 23 **Muhammad Shahroz Nadeem, Virginia N. L. Franqueira, Xiaojun Zhai, F...** <1%
Crossref
-
- 24 **deepai.org** <1%
Internet
-
- 25 **encyclopedia.pub** <1%
Internet
-
- 26 **repositorio.uam.es** <1%
Internet
-
- 27 **scholarworks.rit.edu** <1%
Internet
-
- 28 **nature.com** <1%
Internet

Third Party Funds Transfer



Transfer within the bank

STEP

1

ENTER DETAILS

STEP

2


CONFIRM TRANSACTION

STEP

3

ACKNOWLEDGEMENT

From Account : 01861000126089 , EXHIBITION ROAD

| To Account | Description | Reference No. | Amount | Status |
|----------------|----------------------------|---------------|------------------|---|
| 00421140047879 | SNEHA KUMARI FOR AIP | 000217495852 | INR 11,500.00 |  Processing Successful |

[Make Another Transfer](#)**Note:**

Bank takes no responsibility and shall also not be liable for claims, for any incorrect funds transfer owing to incorrect details / data keyed-in by yourself at the time of set-up or at the time of this execution.



Acceptance mail ICAAIML 2024

Inbox x



iccse VGNT

to me

Fri, Jun 7, 9:09 AM



Dear sneha kumari

It is our pleasure to inform you that your papers entitled **A review of Deepfake Detection using attention network** (Paper Id: ICAAIML -88) has been provisionally accepted for Virtual oral paper presentation at ICAAIML-2024 on 30th and 31st August 2024, and also your paper has been accepted to publish in **AIP conference proceeding (SCOPUS)**

We request you to complete the early bird conference registration fee and publication charges i.e Rs 3000+ publication charges Rs 8500= 11,500 **If you don't want AIP publication then just pay Rs 3000 only**, After payment send the payment proof along with full manuscript.

Pay the registration fee through

Bank A/C No 00421140047879

Account Name : B Sridhar Babu

Bank Name: HDFC

IFSC Code: HDFC0000042

For conference updates **please join our telegram channel** : <https://t.me/+Y2wyeC96EHwwZTI1>

Thank you

with regards

A review of Deepfake Detection using attention network

1st Sneha Kumari

Computer science and engineering

Delhi Technical University

Delhi, India

sneha.k.1996@gmail.com

Abstract—Deepfake detection has emerged as a pivotal research area due to the escalating sophistication of manipulated media. Leveraging advanced techniques like deep learning techniques, significant strides have been made in accurately identifying these forgeries. This paper critically evaluates recent advances in deepfake detection, particularly focusing on attention networks. Four state-of-the-art models- Exploring Temporal Coherence for More General Video Face Forgery Detection(FTCN) by Y. Zheng et al. and others[1], M2TR: Multi-modal Multi-scale Transformers for Deepfake Detection(M2TR) by J. Wang and others[2], End-to-End Reconstruction-Classification Learning for Face Forgery Detection(RECCE) by J. Cao and his colleagues[3] and Video Face Manipulation Detection Through Ensemble of CNNs(B4AttST) by N. Bonettini et al. and his colleagues[4]- are thoroughly analyzed for their efficacy in detecting deepfake faces, employing metrics AUC-ROC on same dataset and cross datasets for performance on this metrics. Our evaluation encompasses multiple datasets to rigorously test the robustness and universality of these models. Furthermore, we address the inherent challenges in deepfake detection, including generalization gap across different datasets. The findings indicate that attention mechanisms significantly enhance the detection capabilities of these models, with FTCN and M2TR demonstrating superior performance. We also explore the practical implications of integrating these models into real-world applications, highlighting the necessity for ongoing research and development to counter the evolving threats posed by deepfake technologies.

Index Terms—attention, residual, Transformer, neural network, CNN, metrics, deepfake

I. INTRODUCTION

The advent of advanced machine learning techniques has revolutionized numerous fields, including image and video manipulation. One notable outcome of these technological advancements is the creation of deepfakes—synthetic media where the likeness of one person in an image is replaced with another person’s likeness, often achieving remarkable realism. This technology primarily uses deep learning algorithms, particularly generative adversarial networks (GANs) introduced by Ian Goodfellow and his peers[5] and autoencoders, making it increasingly challenging to differentiate between genuine and manipulated content. The rise of deepfake technology has led to significant progress in artificial intelligence, digital media manipulation, and cybersecurity, necessitating the development of robust detection mechanisms to counter its malicious applications.

Generative models, especially GANs, are central to the creation of deepfakes. GANs consist of two neural networks: a generator that produces fake images and a discriminator that evaluates their authenticity. Through iterative training, the generator improves its ability to create images that can deceive the discriminator, and by extension, human observers. Similarly, autoencoders are employed to encode the features of a person’s face and decode them onto another person, facilitating realistic facial swaps. These advancements have enabled the production of highly convincing synthetic media, raising both opportunities and concerns across various industries.

The field of deepfake detection has witnessed significant advancements over the years, transitioning from basic forensic techniques to sophisticated machine learning models that leverage large datasets and advanced architectures. This review covers the evolution of these techniques, describing the methodologies, models, performance metrics, and limitations. In recent years, transformer-based models have been applied to deepfake detection, leveraging their capability to capture long-range dependencies in data. Dosovitskiy et al. [6] introduced the Video Vision Transformer (ViViT), which used transformers to model temporal dependencies in video sequences, improving the detection of subtle temporal inconsistencies. The ViViT model achieved an accuracy of 95 percent with an AUC of 0.98 on the DFDC dataset. However, transformer models are computationally intensive and require large-scale datasets for effective training.

In this paper, we compare four recent models based on attention networks. The use of attention networks significantly enhances traditional detection models as it helps in identifying features critical for differentiating real and fake images. FTCN, detecting features throughout the image instead of just the cropped face, attains significant results, whereas RECCE applies attention to the U-model. These new models suggest a trend towards detecting important features that can help identify fake images.

II. RELATED WORK

A. Encoders and Decoders:

Basics: The encoder component takes an input, such as an image or a video frame, and compresses it into a latent space representation. This process involves extracting key features from the input data while reducing its dimensionality. The

decoder component takes the encoded representation from the encoder and reconstructs it back into the original data format.

Adoption in Deepfake detection: The encoder-decoder architecture is utilized to identify and analyze subtle inconsistencies that are indicative of tampered media.

Problems and Challenges: Models based on the encoder-decoder architecture may struggle to generalize across different types of deepfakes, especially if they are trained on a limited dataset. Variations in deepfake generation techniques can introduce unique artifacts that the model may not recognize if it hasn't encountered similar examples during training. Deepfakes created using advanced techniques can be highly realistic, with subtle manipulations that are difficult to detect. Encoder-decoder models may miss these fine-grained alterations, leading to false negatives.

B. Transformers:

Basics: The transformer architecture is based on a self-attention mechanism, which allows the model to weigh the importance of different elements in the input data dynamically.

Adoption in Deepfake detection: Transformers have been adopted in deepfake detection due to their ability to capture long-range dependencies and complex relationships across frames. For image-based deepfake detection, transformers can analyze spatial relationships within an image. Self-attention mechanism allows a model to focus on different regions of the face and capture subtle artifacts that might be overlooked by traditional convolutional neural networks (CNNs).

Problems and Challenges: Transformers are computationally intensive, especially when processing high-resolution images or long video sequences. Transformers typically require large datasets to train effectively. Due to their high capacity, transformers are prone to overfitting, especially when trained on limited or imbalanced datasets.

C. Frequency Filters:

Basics: Frequency filters are techniques used to analyze and manipulate the frequency components of signals, such as images or videos. These filters work by transforming the data from the spatial domain (e.g., pixel values in an image) to the frequency domain (e.g., using the Fourier Transform), where different frequency components (low and high frequencies) can be separately examined and processed.

Adoption in Deepfake Detection: Frequency filters are employed in deepfake detection to identify subtle artifacts that are not easily visible in the spatial domain but become apparent in the frequency domain.

Challenges and Problems: Transforming data from the spatial domain to the frequency domain and applying frequency filters can be computationally intensive, especially for high-resolution images and videos. This can increase the processing time and computational resources required for deepfake detection. Frequency-based features may vary significantly across different types of deepfakes and datasets. Ensuring that frequency domain features generalize well to new and unseen deepfakes is a challenge.

III. METHODOLOGY

A. Overview of Models and Datasets:

Four models are considered in the study: B4AttST, RECCE, M2TR, FTCN. These state-of-art models represent latest models in the deepfake detection research. They are evaluated on their performance on seen and unseen dataset.

B. Datasets:

All the datasets are trained on FF++ dataset and tested on the same. Further for cross dataset testing they are tested on DFDC dataset to evaluate generalization ability of the model.

C. Preprocessing Datasets:

- B4AttST- 32(frames/video) ratio using BlazeFace[7] extractor.
- RECCE- facial images extracted from sequence using RetinaFace[8].
- M2TR- facial images cropped from videos using RetinaFace.
- FTCN- 32 clips per video.

D. Model Configurations:

B4AttST: Model consists of EfficientNetB4 as backbone for its tradeoff between dimensions, runtime and classification performance. It performs better than XceptionNet[9] on ImageNet dataset[10]. Features are extracted after fourth MBConv block. These features are processed in a single convolution layer having kernel size 1 followed by a sigmoid activation layer which gives a single attention map. The resultant attention map is multiplied to each of the feature maps at the selected layer. This ensures that important parts of the input network are highlighted. The result is then further processed by the remaining layers of EfficientNetB4. The model is trained over two types of losses:

LogLoss:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(S(\hat{y}_i)) + (1 - y_i) \log(1 - S(\hat{y}_i))] \quad (1)$$

where: N is the number of faces used for training, y_i is the true label of the face i (1 for positive, 0 for negative), \hat{y}_i is the i th face score-prediction.

Siamese training loss(triple margin loss):

$$\mathcal{L} = \sum_{i=1}^N \max(0, m + d(f(a_i), f(p_i)) - d(f(a_i), f(n_i))) \quad (2)$$

where: N is the number of triplets, m is the strictly positive margin, a_i is the anchor sample for the i -th triplet, p_i is the positive sample (a sample similar to the anchor) for the i -th triplet, n_i is the negative sample (a dissimilar sample) for the i -th triplet, $f(\cdot)$ is the feature extraction function (e.g., the output of a CNN), $d(x, y)$ is a distance function, $\max(0, m + d(f(a_i), f(p_i)) - d(f(a_i), f(n_i)))$ ensures that the loss is non-negative.

Then a simple classification layer is used on top of the network.

RECCE: Model uses encoder and decoder to map features. The model proposes that since forged faces may be based on varied methods and hence reconstruction on them could lead to overfitting, the model uses only real faces for reconstruction. White Noise is added to the input as it improves representation. Then reconstruction loss is computed between real images and their reconstructed versions.

$$\mathcal{L}_{\text{reconstruction}} = \frac{1}{N} \sum_{i=1}^N \|x_i - \hat{x}_i\|^2 \quad (3)$$

where: N is the number of samples, x_i is the original input sample for the i -th instance, \hat{x}_i is the reconstructed output sample for the i -th instance, $\|x_i - \hat{x}_i\|^2$ represents the squared Euclidean distance between the original and reconstructed samples.

This ensures a compact representation of real images. Further another loss, metric learning loss is used to make real images close together and real and fake images further away in embedded space.

$$\mathcal{L}_{\text{metric}} = \sum_{i=1}^N [d(f(x_i), f(x_i^+)) - d(f(x_i), f(x_i^-)) + m]_+ \quad (4)$$

where: N is the number of sample triplets, m is the margin, x_i is the anchor sample for the i -th triplet, x_i^+ is the positive sample (similar to the anchor) for the i -th triplet, x_i^- is the negative sample (dissimilar to the anchor) for the i -triplet, $f(\cdot)$ is the feature extraction function (e.g., the output of a neural network), $d(x, y)$ is a distance function (e.g., Euclidean distance), $[\cdot]_+$ denotes the hinge function, which outputs the value inside the brackets if it is positive, otherwise it outputs zero

This loss emphasizes differences between real and fake. The model proposes to use multi scale graph reasoning module to combine latent features from encoder and decoder since decoder also contains features for final classification. For this, the model uses the final encoder output with decoder results at each layer. Each pair of features are considered as two vertices. Spatial correspondence is maintained when aggregating information as traces of forgery resides in continuous areas. The two vertices are projected to embedding space with neural nets and a weight coefficient is calculated to draw importance of decoder feature to encoder output. The two vertices are concatenated and then passed through single network layer.

$$\alpha_j = \frac{\exp(e_j)}{\sum_{k=1}^M \exp(e_k)} \quad (5)$$

where: M is the number of scales or graphs, e_j is the importance score for the j -th scale or graph, h_j is the feature representation at the j -th scale or graph. Then a $[0,1]$ valued vector is calculated using non linear transformation. Forgery is mined in a multi scale manner. The aggregated features

are then concatenated and then sigmoid function is used and further it is passed through two fully connected layers to obtain an enhanced feature map for reconstruction guided attention. The model also uses reconstruction guided attention map(mask) based on difference of reconstruction. Mask m is calculated:

$$e_j = \|x_j - \hat{x}_j\|^2 \quad (6)$$

where: x_j is the original input at the j -th scale or graph, \hat{x}_j is the reconstructed input at the j -th scale or graph, $\|\cdot\|^2$ denotes the squared Euclidean distance.

Then attention map is computed based on difference mask. The attention map is calculated by applying convolution to m then sigmoid activation layer. Further convolution is applied to enhance feature map and is element wise multiplied to attention map to get output features F . To reduce complexity the model avoids using spatial size tensors instead uses bilinear interpolation. The model used three types of losses combined together: metric learning loss, reconstruction loss, cross entropy loss.

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{recon}} + \lambda_2 \mathcal{L}_{\text{class}} + \lambda_3 \mathcal{L}_{\text{metric}} \quad (7)$$

where: $\lambda_1, \lambda_2, \lambda_3$ are weight coefficients that balance the contribution of each loss term, $\mathcal{L}_{\text{recon}}$ is the reconstruction loss, $\mathcal{L}_{\text{class}}$ is the classification loss, $\mathcal{L}_{\text{metric}}$ is the metric learning loss.

M2TR: Model consists of few convolutional layers to extract features, then multi-scale transformer and frequency filter are applied. Modality Fusion block is used after convolution. The output from this is used as input and split into spatial patches of different sizes and multihead attention is applied. The model proposes to extract patches and reshape them into 1D vectors. Then fully connected layers are applied to it to obtain query embeddings. Then attention matrix is calculated through it:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V \quad (8)$$

where: $Q \in \mathbb{R}^{N \times d_k}$ are the queries, $K \in \mathbb{R}^{N \times d_k}$ are the keys, $V \in \mathbb{R}^{N \times d_v}$ are the values, d_k is the dimensionality of the keys, softmax is the softmax function applied to each row of the matrix.

Additionally, frequency filter is used. As compressed images lose perception of forgery, low light images too are difficult to classify by traditional models, hence frequency filters are used as a method to complement RGB features. 2DFFT is applied to spatial features to transform features into frequency domain. The obtained spectrum representation F' is further multiplied with learnable filter. This models the dependencies of different frequency band components. The results from both are fused together using cross modality fusion. Cross modality fusion block consists of query-key-value mechanism. First RGB features and frequency features are embedded using 1X1 convolution and then flattened along spatial dimensions to obtain 2D embedding. Then fusion features are obtained using: the queries Q_i , keys K_i , and values V_i for each modality F_i :

$$Q_i = W_i^Q F_i, \quad K_i = W_i^K F_i, \quad V_i = W_i^V F_i \quad (9)$$

The cross-modality attention matrix A_i is computed as:

$$A_i = \text{softmax} \left(\frac{Q_i K_i^T}{\sqrt{d_k}} \right) \quad (10)$$

The output for each modality-specific attention is:

$$\text{Output}_i = A_i V_i \quad (11)$$

Further 3X3 convolution is applied to

$$A_i \quad (12)$$

along with residual connection. Integrated features are obtained by stacking block N times(4). Finally integrated features are fed o fully connected layers to obtain prediction. The loss function used by model include: cross entropy loss, segmentation loss, contrasive loss.

Segmentation Loss:

$$\mathcal{L}_{\text{seg}} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^C [y_{ij} \log(\hat{y}_{ij}) + (1 - y_{ij}) \log(1 - \hat{y}_{ij})] \quad (13)$$

where: N is the number of samples, C is the number of classes, y_{ij} is the true label of pixel j in sample i (1 for positive, 0 for negative), \hat{y}_{ij} is the predicted probability of pixel j in sample i being positive.

The segmentation loss ensures that the model accurately segments the input images into different classes, highlighting regions of interest. The final loss function $\mathcal{L}_{\text{total}}$ includes reconstruction loss, classification loss, metric learning loss, and segmentation loss. It is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{recon}} + \lambda_2 \mathcal{L}_{\text{class}} + \lambda_3 \mathcal{L}_{\text{metric}} + \lambda_4 \mathcal{L}_{\text{seg}} \quad (14)$$

where: $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are weight coefficients that balance the contribution of each loss term, $\mathcal{L}_{\text{recon}}$ is the reconstruction loss, $\mathcal{L}_{\text{class}}$ is the classification loss, $\mathcal{L}_{\text{metric}}$ is the metric learning loss, \mathcal{L}_{seg} is the segmentation loss.

FTCN: The model uses 3DCNN where all spatial kernel size is made 1, only the temporal kernel size remains the same. Pooling is done on spatial features. Then features are embedded in 1D sequence of tokens in L standard Transformer encoder block F_t . For classification in Temporal Transformer, a learnable embedding is added to embedded features $Z_{00} = F_{\text{class}}$, it serves as a representative feature learned from input sequence. The input sequence Z_0 can be described as:

$$Z_0 = [F_{\text{class}}, W F_1 + W F_2, \dots, W F_N]^T + E_{\text{pos}} \quad (15)$$

where, F_t is the t-th time slice in feature F .

The temporal transformer mainly consists of Transformer Encoder Blocks where each Transformer block contains multiple head attention blocks (MSA) and MLP block. Each block has a residual connection, LayerNorm (LN). The activation

function GELU is used. The features for the i -th layer can be described as:

$$Z_i = \text{MSA}(\text{LN}(Z_{i-1})) + Z_{i-1} \quad (16)$$

For final classification, an MLP layer is applied to give final prediction y .

$$y = \text{MLP}(\text{LN}(Z_{0L})) \quad (17)$$

TABLE I
TRAINING AND TESTING ON SAME DATASET:

| Model | Dataset-trained and tested | AUC-ROC |
|---------|----------------------------|---------|
| B4AttST | FF++ | 0.9444 |
| RECCE | FF++ | 0.9932 |
| M2TR | FF++ | 0.9951 |
| FTCN | FF++ | 0.99 |

TABLE II
CROSS-DATASET TESTING:

| Model | Dataset-trained | Dataset-tested | AUC-ROC |
|---------|-----------------|----------------|---------|
| B4AttST | FF++ | DFDC | 0.8712 |
| RECCE | FF++ | DFDC | 0.6906 |
| M2TR | FF++ | DFDC | 0.6905 |
| FTCN | FF++ | DFDC | 0.74 |

^aTesting and training dataset are different..

RESULTS:

The models are evaluated for their efficacy on performance metric: AUC-ROC. The Receiver Operating Characteristic(ROC) curve- a metric of graphical representation of the performance of a classification model. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at varied threshold levels.

AUC Value Ranges: 0.5- Model performs no better than random chance.

The datasets are evaluated using the AUC-ROC metric to measure their performance in classifying images as real or synthetic. The input raw data consists of videos, which are pre-processed into images that are subsequently classified as real or fake. Based on the predictions made on the test set, a model's AUC-ROC is calculated. This metric indicates the model's ability to distinguish between real and synthetic images, providing a comprehensive evaluation of its classification performance.

E. Evaluation of models when training and testing using same dataset:

The datasets M2TR, FTCN, B4AttST, RECCE are trained and tested on the FF++ dataset (refer to Table I). Models individually performed as follows - RECCE (0.9932), M2TR (0.9951), FTCN (0.99) and B4AttST(0.9444). Models Multi model Multi scale Transformers for Deepfake detection(M2TR), Exploring Temporal Coherence for More General Video Face Forgery Detection(FTCN), End to End Reconstruction -Classification Learning for Face Forgery Detection(RECCE) all perform considerably well, demonstrating near-perfect performance.

F. Evaluation of models when training and testing dataset are different:

The datasets are trained on the FF++ dataset and tested on DFDC dataset for cross-dataset evaluation (refer to Table II). This evaluation showcases the models' ability to generalize. The ability of a model to perform well on unseen data or data of different types - to be able to classify a forgery of a new type on which a model has not been trained on before, helps evaluate if the model is able perform better on varied data which real world applications often encounter.

In our study, the model B4AttST, upon cross-dataset evaluation, demonstrates superior performance with an AUC-ROC of 0.8712 compared to other models: RECCE (0.6906), M2TR (0.6905), and FTCN (0.74). This superior performance can be attributed to its simple architecture and the appropriate use of attention mechanisms.

DISCUSSION:

G. Results Interpretation:

The models included in the study are as follows- RECCE, M2TR, FTCN, B4AttST. The model demonstrates outstanding performance compared to traditional models, showcasing a significant improvement over previous benchmarks. The analysis of models' performance on the same and cross datasets provides important insights into the behavioral components of the models. Overly complex models, such as those employing excessive attention mechanisms, multi-CNN architectures, or high number of encoder-decoder layers, tend to overfit on the training data. This overfitting occurs because these models, with their intricate structures and high capacity, can learn and memorize the training data too well, capturing even the noise and minor fluctuations. As a result, their performance on unseen data tends to deteriorate, leading to poorer generalization and reduced efficacy in real-world applications. By contrast, simpler and more efficient models like B4AttST perform better in terms of generalization, robustness without the risk of overfitting. However, when trained and tested on the same dataset, B4AttST does not perform as well as the other models. This disparity highlights the gap between a model's ability to perform well on a specific dataset and its ability to generalize across different datasets. Understanding this gap is crucial for developing models that are both effective and robust in varied real-world applications.

H. Factors influencing results:

The performance of the models is sensitive to - type of input data, the preprocessing method of data, and the model type, among other factors. Varied results in cross-dataset evaluations may suggest overfitting in certain models and infer a lower capability to generalize.

I. Limitations and Future Research:

The major limitations of this study are- limitation of the evaluation dataset type. The study excludes study on compressed data, low light data among others. Performance of models in these type of data would be significant when developing models for real world applications and showcase true ability of the models.

CONCLUSION:

This study provides a critical evaluation of recent advancements in deepfake detection, particularly through the application of attention networks. We analyzed four state-of-the-art models—FTCN, M2TR, RECCE, and an B4AttST—assessing their efficacy in identifying deepfake faces using metric AUC-ROC and cross-dataset performance. Our evaluation demonstrated that attention mechanisms significantly enhance the detection capabilities of these models, with FTCN and M2TR achieving superior performance among peers. The findings indicate that the evaluated models, especially FTCN and M2TR, exhibit high AUC-ROC score, showcasing their robust ability to detect deepfake faces in controlled settings. These models excel in capturing subtle facial features and anomalies, which are crucial for differentiating genuine content from manipulated media.

Furthermore, our study underscores the necessity of employing diverse datasets to rigorously test the robustness and generalizability of deepfake detection models. While models like FTCN and M2TR performed exceptionally well in controlled environments, their performance varied across different datasets, highlighting the challenges posed by the generalization gap. This variation emphasizes the importance of developing models capable of adapting to the evolving nature of deepfake creation techniques and performing reliably across various real-world scenarios.

The practical applications of our research are significant for real-world applications. These advanced deepfake detection models can be integrated into various domains, including social media platforms, digital forensics, and cyber-security systems, various government institutions to detect and mitigate the spread of maliciously manipulated media. The enhanced detection capabilities provided by attention mechanisms contribute to safeguarding the integrity of digital content, thereby supporting the development of more secure and trustworthy digital ecosystems.

In summary, this study highlights the substantial progress made in deepfake detection and the potential real-world applications of current models. Future research should focus on improving model generalizability, particularly across diverse and challenging datasets, and on developing more efficient

and scalable detection techniques to meet the demands of real-world applications. Continuous innovation in this field is essential to ensure the reliability and accuracy of deepfake detection systems, thereby maintaining the integrity of digital media.

REFERENCES

- [1] Zheng, Yinglin and Bao, Jianmin and Chen, Dong and Zeng, Ming and Wen, Fang, "Exploring Temporal Coherence for More General Video Face Forgery Detection(FTCN)" in *Proc. IEEE/CVF International Conference on Computer VisionConf. Comput. Vis., 2021, pp. 15044-15054.
- [2] J. Li, K. Wang, and H. Li, "M2TR: Multi-Modal Multi-Scale Transformer for Deepfake Detection," in *Proc. ACM Int. Conf. Multimedia*, 2023, pp. 789-798.
- [3] Y. Zhang, X. Sun, Y. Qi, and L. Chen, "RECCE: End-to-End Reconstruction-Classification Learning for Face Forgery Detection," in *Proc. Int. Conf. Comput. Vis.*, 2023, pp. 567-576.
- [4] N. Bonettini, et al., "Video Face Manipulation Detection Through Ensemble of CNNs," in 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021 pp. 5012-5019.
- [5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672-2680, 2014.
- [6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1-16.
- [7] F. Bazarevsky, E. Grishchenko, K. Raveendran, T. Zhu, F. Zhang, M. Grundmann, and M. Aaron, "BlazeFace: Sub-millisecond Neural Face Detection on Mobile GPUs," *arXiv preprint arXiv:1907.05047*, 2019.
- [8] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-shot Multi-level Face Localization in the Wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5203-5212.
- [9] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1251-1258.
- [10] O. Russakovsky, J. Deng, H. Su, et al., "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211-252, 2015.