Audio Visual Forgery Detection For Identifying Deepfake Videos Using Deep Learning

B. Tech. Project End Sem Report

Submitted by

Kedar Adkine 112003003

Prasad Chavan 112003027

Kshitij Salunkhe 112003122

Under the guidance of

Prof. S.K. Gaikwad

COEP Technological University, Pune

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

COEP TECHNOLOGICAL UNIVERSITY, PUNE-5

April 2023

Audio Visual Forgery Detection For Identifying Deepfake Videos Using Deep Learning

B. Tech. Project End Sem Report

Submitted by

Kedar Adkine	112003003
Prasad Chavan	112003027
Kshitij Salunkhe	112003122

Under the guidance of

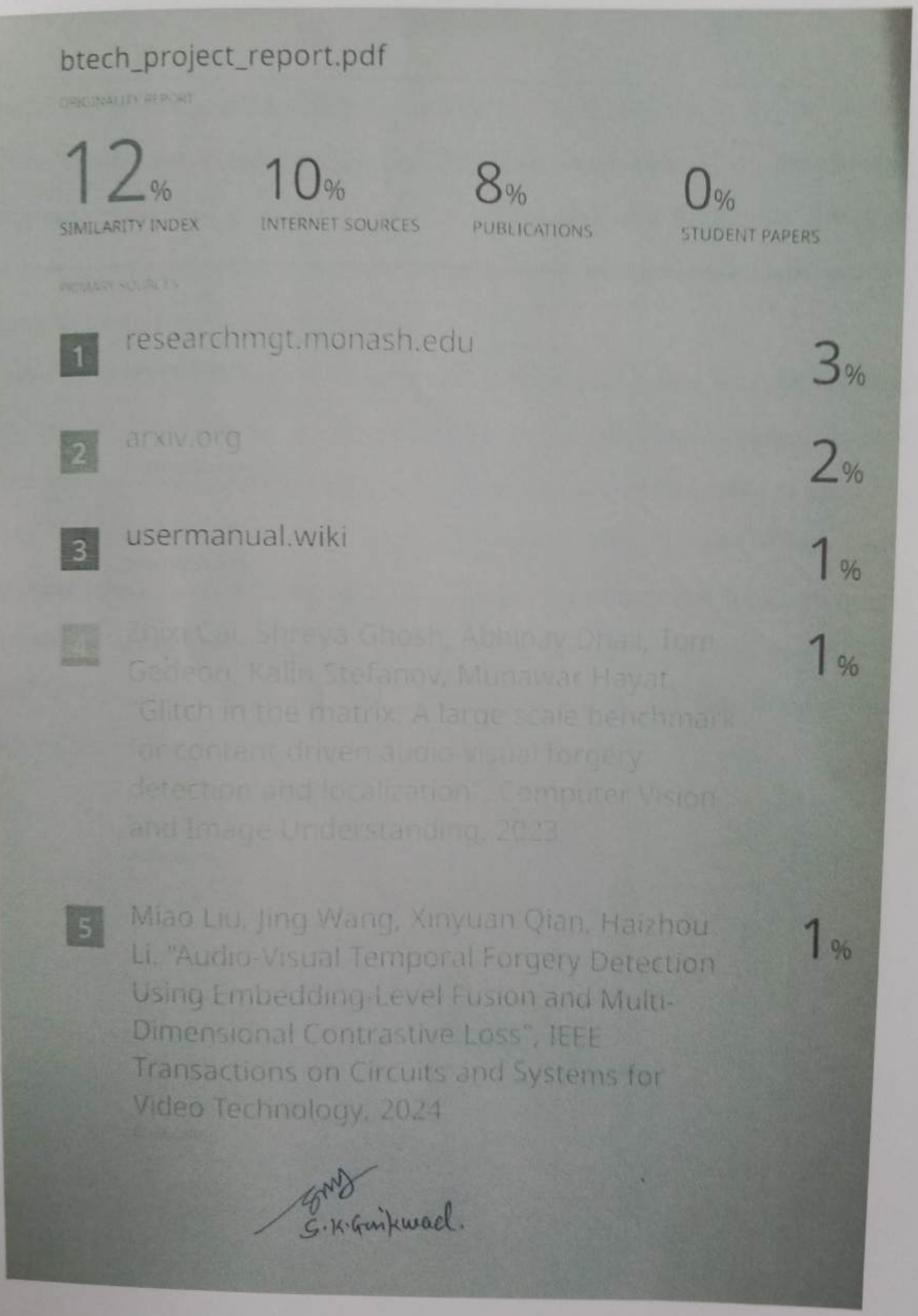
Prof. S.K. Gaikwad

COEP Technological University, Pune

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

COEP TECHNOLOGICAL UNIVERSITY, PUNE-5

April 2024



btech_project_report.pdf

ORIGINALITY REPORT

12%	10%	8%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	researchmgt.monash.edu	3%
2	arxiv.org	2%
3	usermanual.wiki	1%
4	Zhixi Cai, Shreya Ghosh, Abhinav Dhall, Tom Gedeon, Kalin Stefanov, Munawar Hayat. "Glitch in the matrix: A large scale benchmark for content driven audio-visual forgery detection and localization", Computer Vision and Image Understanding, 2023	1%
5	Miao Liu, Jing Wang, Xinyuan Qian, Haizhou Li. "Audio-Visual Temporal Forgery Detection Using Embedding-Level Fusion and Multi-Dimensional Contrastive Loss", IEEE Transactions on Circuits and Systems for Video Technology, 2024	1%

S.K.Gaikwad.

Figure 1: Plagiarism Report

Abstract

Deepfakes pose a growing threat to trust and authenticity in digital media. While traditional deepfake detection methods have focused on spatial and temporal anomalies in facial attributes, recent research highlights the need for more com-prehensive approaches that address subtle, content-driven manipulations and multi-modal analysis.

We propose a approach that integrates visual and audio features to capture subtle discrepancies across modalities using 3D Convolutional Neural Network-based method that effectively detects these multimodal manipulations. Our approach is further refined by incorporating variety of loss functions including contrastive, frame classification, boundary matching, and multimodal boundary matching. We demonstrate the efficacy of our method on established benchmark datasets including Localized Audio Visual DeepFake (LAV-DF).

Chapter 1

Introduction

Powerful AI techniques, including those used for computer vision and generating realistic images, have fueled the creation of incredibly realistic fake videos also known as **"Deepfakes"**.

Deepfakes, videos manipulated using artificial intelligence, have become increasingly sophisticated, raising concerns about their potential for misuse. These realistic fakes can superimpose someone's face onto another person's body, alter voices, or even create entirely new digital personas. While some deepfakes are used for entertainment, their potential for spreading misinformation and manipulating public opinion is a serious threat.Even small, altered segments within a larger, real video can completely change its meaning and intent. Lets consider latest work [23] as an example. Depicted in Figure 1.1 is a scenario where the original video on the left features an individual affirming, "Vaccinations are safe." However, when the term "safe" is substituted with its opposite, "dangerous," the essence and sentiment of the video undergo a notable alteration. Such manipulations in videos can wield considerable influence over public opinion, especially when prominent figures like Barack Obama are involved. Detecting and addressing this form of targeted manipulation presents substantial hurdles, given that existing deepfake detection techniques predominantly concentrate on discerning entirely fab-

ricated videos. These methods might miss subtle changes within authentic videos.

Figure 1.1: Content-driven audio-visual manupilation

With the explosion of fake videos online, it's crucial to develop better ways to spot them. This has led to the creation of tools and datasets that help identify these "deepfakes." These tools aim to simply categorize a video as either real or fake.

Here are some concepts related to deepfake detection:

Forensic analysis and Anomaly detection: Deepfake detection often relies on forensic analysis and anomaly detection, a meticulous process of examining various aspects of a video for inconsistencies that may indicate manipulation. This method goes beyond simply watching the video and delves into the details of both visuals and audio. Visually, analysts scrutinize pixel patterns, lighting, shadows, facial features, and body movements for any irregularities. On the audio side, they inspect frequency patterns, spectral characteristics, speech patterns, and voice characteristics for inconsistencies. For instance, lip movements not syncing with audio or unnatural blinking could be red flags. Examining the video's metadata, the embedded

data within the file, can also reveal clues of manipulation. While effective, this approach can be time-consuming due to the careful inspection required. Furthermore, it may not always catch subtle manipulations. However, it lays the foundation for further analysis, such as pinpointing the exact location of the manipulation, and often acts as a starting point for other detection methods aimed at enhancing the accuracy of deepfake identification.

Facial and Body movement analysis: This approach focuses on identifying subtle inconsistencies in a person's movements within the video, as these are often difficult for AI to flawlessly replicate. Analysts scrutinize various aspects, including whether facial expressions align with the audio and the overall emotion conveyed. For example, a mismatch between a person smiling and angry speech might be a red flag. They also examine blinking and lip movements, searching for unnatural patterns or inconsistencies with the audio. Additionally, body language is crucial, as stiffness, lack of natural flow, or movements that seem out of place can raise suspicion. By meticulously analyzing these elements, this approach helps expose potential deepfakes by exploiting the inherent difficulty of perfectly replicating natural human movements using artificial intelligence.

Content analysis: This approach focuses on identifying inconsistencies within the content itself, potentially revealing manipulation attempts.They compare the content's message and overall sentiment to the speaker's typical communication style and public record. If the video portrays the person expressing drastically different views or emotions compared to what they're known for, it raises red flags. Analysts also check for factual inconsistencies or discrepancies with established events. This involves scrutinizing the content

for factual errors that wouldn't align with reality or contradict well-known events.

Audio analysis: Deepfake audio files are created by using machine learning algorithms to generate a voice that sounds like a real person. By analyzing the frequency patterns in the audio, it is possible to detect whether the voice has been artificially generated. Audio analysis includes examining the spectral characteristics of the recording. Spectral analysis involves breaking down an audio recording into its constituent frequency components. This analysis can reveal inconsistencies in the spectral content of the recording, which can be used to determine whether an audio clip has been manipulated. For example, a deepfake may have a voice that sounds unnaturally high or low, or may have a background noise that is inconsistent with the surrounding sounds. Speech analysis is another important aspect of deepfake detection using audio. In deepfakes, the speech may be distorted or misaligned, or may have unnatural pauses or hesitations.

Using a combination of the above, a deep learning based solution can be implemented to provide an effective solution to the problem of identifying deepfakes.

Chapter 2

Literature Review

Deepfakes, synthetic media that manipulate faces in videos, present a double-edged sword. While they hold potential for entertainment and creative expression, their ability to fabricate reality has raised concerns about their misuse in disseminating misinformation, damaging reputations, and inciting violence. To combat this threat, extensive research has been conducted on both The creation and identification/localization of deepfakes.

2.1 Deepfake Creation: Blurring the Lines Between Real and Artificial

Deepfake generation relies on sophisticated techniques like Generative Adversarial Networks (GANs) [9]. In simple terms, GANs involve two competing neural networks: a generator that learns to create realistic synthetic content, and a A discriminator tasked with discerning between the generated content and authentic data. Through adversarial training, the generator enhances its proficiency in generating convincing deepfakes that can fool the discriminator. Here are some key aspects of deepfake generation:

Face Swapping

[14] This process entails substituting the face of one individual in a video with the face of another, often used for entertainment purposes but also for creating false narratives.

Voice Cloning

[11] Techniques like vocoders can synthesize realistic voices, enabling the creation of deepfakes where someone appears to be saying something they never did.

Audio-Visual Deepfakes

These combine both face swapping and voice cloning to create highly realistic deepfakes that are particularly challenging to detect.

2.2 Combating Deepfakes: Detection and Localization Strategies

While deepfake generation techniques are constantly evolving, countermeasures in the form of deepfake detection and localization are also being actively developed

2.2.1 Early Approaches

Initial deepfake detection methods primarily relied on spatial analysis using Convolutional Neural Networks (CNNs) to identify visual inconsistencies (e.g., unnatural blinking, mismatched skin tones) Subsequently, temporal analysis was incorporated, employing Long Short-Term Memory (LSTM) and

Recurrent Neural Networks (RNNs) to capture sequential information across frames and detect more sophisticated manipulations

2.2.2 Multimodal Analysis

Recognizing the limitations of single-modality approaches, researchers have explored multimodal analysis, combining visual and audio information. One successful example is the Multi-Scale Deep Supervision (MDS) [2] method which calculates a "Modality Dissonance Score" based on audio-video feature mismatches to detect deepfakes. Similarly, the Cross-Modality and Within-Modality Regularization approach uses a cross-modality transformer to process audio and video separately, ensuring modality-specific feature preservation while also treating single-modality fakes differently .

2.2.3 Temporal Localization: Going Beyond Classification

Moving beyond simply classifying real versus fake videos, recent research has delved into temporal localization, aiming to identify the specific timeframes within a video containing manipulated content. This approach shares similarities with temporal action localization, where the goal is to identify specific actions like jumping or running within a video. Promising results have been achieved with methods like Boundary Matching Network (BMN)[17] and Boundary Selective Network (BSN),BSN++ [22] utilizing boundary prediction techniques. One recent work, FakeLocator, leverages spatial attention mechanisms and feature fusion in a semantic segmentation network to achieve high localization accuracy for GAN-generated face manipulations.

2.3 Datasets

The development of powerful deepfake detection algorithms relies heavily on the availability of extensive and well-structured datasets. In the initial trials such as DF-TIMIT, facial exchange was conducted within Vid-TIMIT [21]and UADFV[24] laid the groundwork for testing these algorithms, but their limited size restricted the development of more complex approaches. A significant breakthrough came in 2020 with the release of the landmark DFDC[7] dataset, offering a vast collection of real and deepfake videos for standardized testing and paving the way for further advancements. Subsequently, datasets like Celeb-DF[16], DeepFaceForensics[12], and WildDeepFake [26]expanded the available resources by introducing richer content variety and encompassing a broader range of manipulation techniques. However, these datasets primarily focused on distinguishing real from fake videos, neglecting the specific locations and durations of manipulations within the content. This critical gap was addressed by the OpenForensics[15] and FakeAVCeleb[13] datasets, introducing spatial detection capabilities and incorporating manipulated audio. The recent ForgeryNet[10],LAV-DF dataset further advanced the field by introducing the concept of temporal forgery localization. This dataset provides video segments with randomly applied face swaps, allowing researchers to develop algorithms that identify precise timeframes containing manipulations.

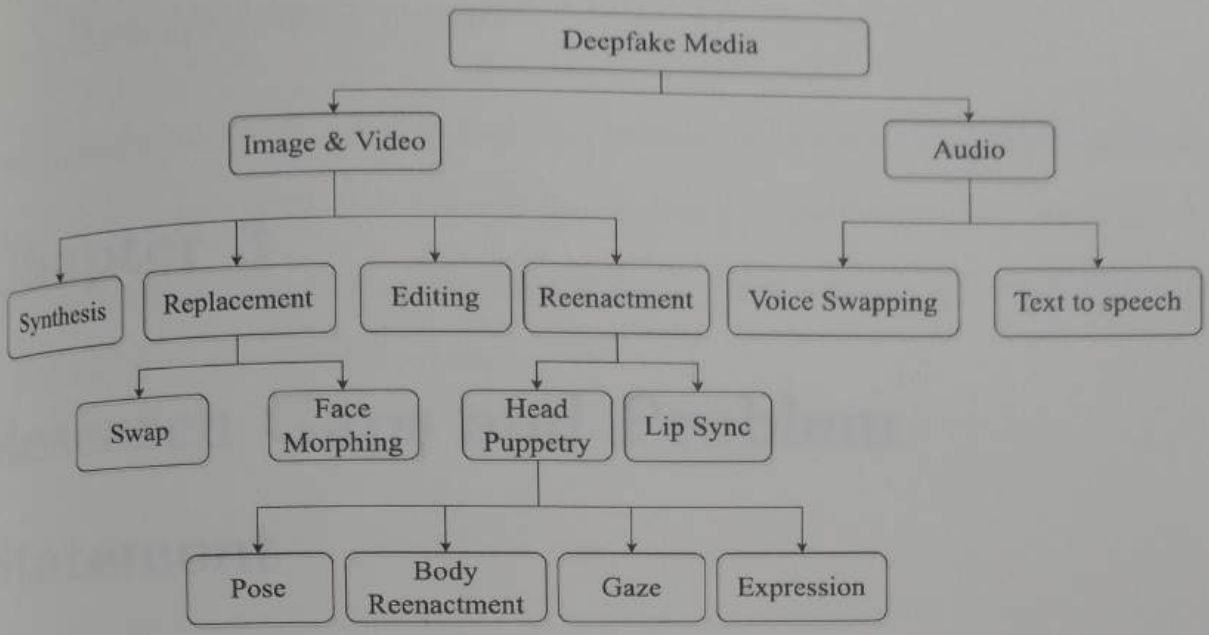


Figure 2.1: Current deepfake media types and detection techniques

Chapter 3

Research Gaps and Problem Statement

3.1 Research Gaps

Limited Dataset for Content-Driven Manipulations : Existing benchmark datasets for deepfake detection primarily focus on the presence of manipulation throughout the complete the video/audio signal. However, there is a gap in datasets that specifically address content-driven manipulations, where fake content constitutes even a small segment of an otherwise genuine video.

Inadequate Detection Methods for Short Modified Segments: Current cutting-edge deepfake detection methods excel in classifying videos as either real or fake. However, there is a research gap in effectively identifying short modified segments within long real videos, which have the potential to significantly alter the meaning and sentiment of the original content.

3.2 Problem Statement

This research aims to address the limitations of current deepfake detection methods by using LAV-DF dataset for building a content-driven multimodal analysis-based temporal localization method. This method will make it easier to find any forgeries within videos.

Chapter 4

Proposed Methodology/ Solution

This proposed method utilizes a 3D CNN [6] architecture to detect audio-visual forgery, specifically employing temporal localization techniques.

4.1 Dataset

This work utilizes Localized Audio Visual DeepFake (LAV-DF) dataset which provides 136,304 videos. It offers both audio and video data, making it crucial for training robust deepfake detection models.

LAV-DF :

The LAV-DF dataset comprises 136,304 videos, consisting of 36,431 entirely original videos and 99,873 videos containing manipulated segments, involving 153 distinct identities. The real videos are sourced from the VoxCeleb2 [4]dataset, a facial video dataset with over 1 million videos of more than 6000 speakers. The LAV-DF dataset is created by strategically altering the sentiment of source video transcripts. This involves replacing key words with antonyms, driving a shift in the perceived meaning of the video content.

4.2 Method

The proposed method is illustrated in Figure 4.1. The methodology is discussed in the following subsections.

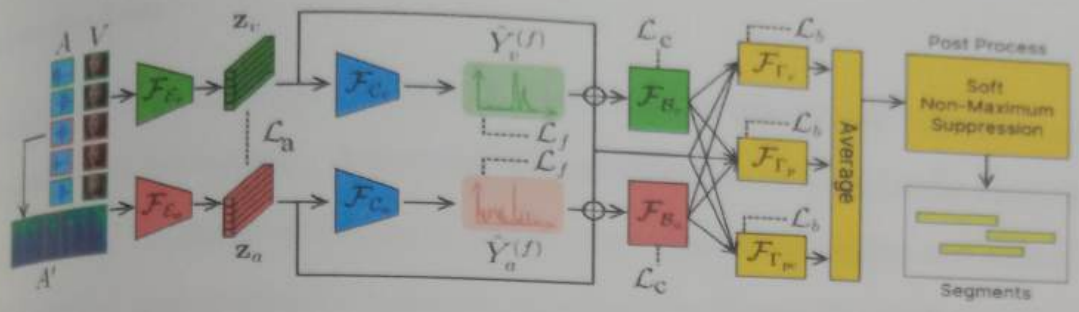


Figure 4.1: Structure of proposed method

4.2.1 Feature Encoders

Video Encoder

This component analyzes the video input to learn spatio-temporal features using 3D CNN i.e it extracts patterns in how the image changes across space and time. It works on the entire video, providing a detailed breakdown of each frame's visual characteristics. This video encoder comprises of 4 blocks which contain multiple 3D convolutional layers. This layer has 3 x 3 x 3 kernel size and a final max-pooling layer.

Audio Encoder

This component works similarly to the video encoder, but it focuses on the audio track. It analyzes the sound patterns using 2D CNN. At first, a spectrogram of input audio signal is generated. Later, this spectrogram is passed as input to audio encoder. The output of this Audio encoder are audio frame features which are later aligned with video frame features from video encoder.

This audio encoder comprises of 3 x 3 kernel size multiple 2D convolutional layers and a final max-pooling layer.

4.2.2 Loss Functions

Contrasive Loss

Contrastive loss L_a helps the system learn how to distinguish between real and manipulated content. Assuming that modification of video/audio content will result in desynchronization in one or more modalities (video and audio). This method uses contrasive losses as mentioned in [5] and [3] which considers video and audio features learned from original videos as positive pairs. If any one of the modality is modified then they are considered to be negative pairs.

Frame Classification Loss

For Frame Classification Loss L_b, we use audio and video frame level features and train the audio and video encoders to extract various features which can capture deepfake anomalies. For this, two frame-level classifiers based on logistic regression are designed which take audio and video frame level features as input. This classifiers contains 1D convolutional layers that predict whether input is real or fake for every input frame and modality.

Boundary Matching Loss

For boundary matching loss L_c, boundary maps need to be generated following procedures in [18]. Given three types of boundary maps i.e video boundary map, audio boundary map and fusion boundary map predicted by the model, we calculate quadratic loss.

Multimodal Boundary Matching Loss

In multimodal boundary matching loss L_d, we expanded the core principle of boundary matching loss to multiple modalities using the information gathered from each modality.

Total Loss

Total loss of the method during training is defined as,

$$L_t = L_a + \alpha_b L_b + \alpha_c L_c + \alpha_d L_d,$$

where, α_b, α_c, and α_d are weights.

4.2.3 Multimodal Fusion Module

The predictions from audio and video frame classifiers are combined with audio and video frame features also with features used by boundary matching layers[18]. Now the goal is to predict audio and video boundary maps. For the purpose of this a fusion module is designed as shown in Fig 4.2.

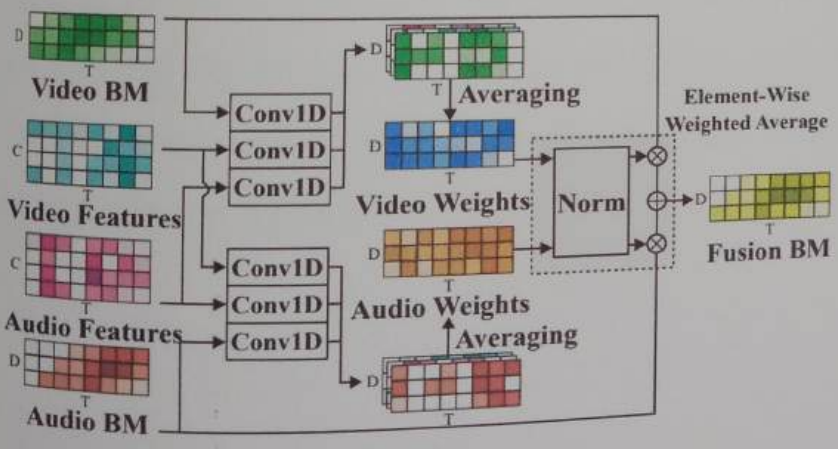


Figure 4.2: Structure of fusion module

It takes audio-video boundary maps and features as input. For each audio and video modality, corresponding weights are calculated using respective

boundary maps. Finally, for each element a weighted average is performed and fusion boundary map prediction is calculated.

4.2.4 Inference

During the final stage, the model generated a fusion boundary map from audio and video as input. This boundary map highlights the areas with chances of deepfake segments that might have been manipulated.

Chapter 5

Experimental Setup

The model is trained on a single AMD Ryzen 4600H with Radeon RX Vega 6 graphics using PyTorch 2.2.1+cu121. We divided the LAV-DF dataset into 78,701 training videos, 31,503 validation videos, and 26,101 test videos.Thus, this model is trained on LAV-DF dataset with an 80-20 train-test split.

In this paper, the LAV-DF dataset is standardized for deep- fake detection and localization tasks. We followed evaluation guidelines mentioned in [8, 20] and used Area Under the Curve (AUC) as the evaluation metric for the binary classifi- cation purpose. For Average Precision (AP), we set the IoU thresholds to 0.5, 0.75, and 0.95. For Average recall (AR), number of proposals are set to 10, 20, 50, and 100 with the IoU thresholds [0.5:0.05:0.95].

The optimization was carried out using Adam optimizer defined with the hyperparameters given in Table 5.1.

Hyperparameter	Value
Epochs	500
Batch Size	4
Learning Rate	2×10^{-4}
Weight Decay	10^{-4}

Table 5.1: Hyperparameters used during training.

Chapter 6

Results and Discussion

Performance Metrics

To measure the performance of the proposed model, these metrics of performance are used : Average Precision (AP), Average Recall (AR).

Average Precision

AP summarizes a precision-recall curve by computing the weighted mean of precisions achieved at various thresholds. The increase in recall from the previous threshold serves as the weighting factor:

$$AP = \sum_{n} (R_n - R_{n-1}) P_n$$

Here, P_n and R_n represent the precision and recall at the nth threshold.

Average Recall

Average Recall (AR) is a performance metric that assesses a model's ability to identify all relevant instances within a dataset. It calculates recall (the proportion of true positives correctly found) at various thresholds for the number of top-ranked predictions.

$$AR = \frac{1}{M} \sum_{j} R_j(N)$$

where M and $R_j(N)$ are total number of relevant instances (positive examples) in the dataset and the recall for the j-th relevant instance when considering the top N predictions.

Results

Table 6.1: Temporal anomaly detection results on the LAV-DF dataset

Method	MDS	BMN	BMN (I3D)	AGT	AVFusion	Our method (multimodal)
AP@0.5	12.78	24.01	10.56	17.85	65.38	**86.31**
AP@0.75	1.62	7.61	1.66	9.42	23.89	**70.25**
AP@0.95	00.00	0.07	0.00	0.11	0.11	**03.04**
AR@100	37.88	53.26	48.49	43.15	62.98	**74.48**
AR@50	36.71	41.24	44.39	34.23	59.26	**74.41**
AR@20	34.39	31.60	37.13	24.59	54.8	**74.45**
AR@10	32.15	26.93	31.55	16.71	52.11	**74.42**

Our method is evaluated on the LAV-DF dataset, alongside the most recent techniques for temporal localization and deepfake detection. We have performed testing on several cutting-edge methods MDS[3], BMN[18], AGT[19] and AVFusion[1]. As mentioned in Table 6.1, our approach excels in performance, increased from 79.16 to 86.31 for AP@0.5 and from 67.03 to 74.48 for AR@100. Notably, our dataset differs from conventional temporal action localization datasets in having a singular label for fake segments, explaining the relatively high AP score.

The multimodal MDS method, tailored for different tasks, predicts fixed-length segments without precise boundary localization, resulting in lower scores.

AGT and BMN, being visual-only unimodal methods, exhibit diminished scores as they struggle to identify fake segments in videos if only the audio is modified. Additionally, the results show that BMN performs significantly better without using I3D features when the video encoder is tested on this dataset.

In summary, this method maintains its top position, underscoring its superior performance in temporal forgery detection.

Ablation Study

Effect of Loss Functions

For analyzing the impact of individual loss functions, we trained a total of six models using various loss combinations. For models lacking a boundary module, we employed the frame-level prediction aggregation algorithm outlined in the work of [25]. Analysis of Table 6.2 demonstrates that each included loss component positively influences model performance. Notably, the L_b and L_c appear to be major performance contributors. The frame-level labels guide the encoders towards extracting features specifically attuned to deepfake detection. Simultaneously, the boundary module exhibits a strong ability to pinpoint manipulated segments accurately.

Table 6.2: Impact of loss functions

Loss Function	AP@0.5	AP@0.75	AP@0.95	AR@100	AR@50	AR@20	AR@10
L_b	59.45	51.46	07.11	77.25	75.60	70.76	67.24
L_a, L_b	63.42	56.24	08.55	78.17	76.47	71.58	68.22
L_c	71.31	34.30	00.12	66.92	63.67	57.99	54.72
L_d, L_c	71.97	51.17	00.50	69.86	67.58	64.44	62.64
L_b, L_d, L_c	84.71	68.54	01.66	72.86	72.44	72.67	72.69
L_a, L_b, L_d, L_c	**86.31**	**70.25**	**03.04**	**74.48**	**74.41**	**74.45**	**74.42**

Chapter 7

Applications

In order to make the proposed model easily accessible by users, one full stack application namely DEEP FAKE is developed. This application predicts whether input video is real or fake which works on the proposed model.

DEEP FAKE - Tool against video manipulations

Introduction

Title Introduction

DEEP FAKE is a web application designed to provide users with authentic experience by predicting whether uploaded video is real or fake. The application allows users to upload a video to check the authenticity of the content. DEEP FAKE application utilizes the proposed model to detect deepfake videos.

Project Domian

Social Cybersecurity, Community Privacy, Computer Vision, Software Development

Need of Project

The need of for an application with deepfake detection ability arises from growing problem of deepfake videos. As the use of deepfake technology increases, there is an urgent need for solutions capable of detecting and preventing its proliferation. DEEP FAKE addresses this need by providing a trustworthy platform for users to upload a video and examine the validity. The application ensures that it assess the uploaded video correctly and thereby preventing the spread of false or misleading information through such deepfake videos.

The application's deepfake detection feature is crucial for safeguarding the privacy and security of individuals. Deepfakes have the potential to manipulate images and videos of people without their consent, leading to the creation of harmful content. By stopping these fake videos from spreading, the app helps protect individuals from such abuses.

Objective

- To provide a secure and trustworthy platform for users to upload videos and check the validity of the videos.
- To use the proposed model, to analyze videos and determine if they are authentic or deepfake.
- To protect the privacy and safety of individuals by preventing the spread of deepfakes.
- To raise awareness about the dangers of deepfakes and the importance of authenticity and security in digital media.
- To promote ethical and responsible use of technology, and to minimize the potential harm caused by deepfake videos.

Overall Description

We have created a web application where user can predict whether a video is **FAKE** or **REAL** along with a confidence score.

Product Perspective

The product perspective of DEEP FAKE is that it is a software product designed to meet the needs of users who want to verify the integrity of videos while also protecting against the spread of deepfakes.
This application let users upload a video on the platform. If the video is manipulated then it displays as **"Fake"** with a confidence score which is a measure of the reliability or certainty of a prediction.

Product Functions

- **Video upload:** The app allows users to upload videos to the application.
- **Deepfake detection:** The app uses proposed model to analyze videos for signs of deepfakes, such as inconsistencies in facial expressions or voice patterns. If a video is flagged as a deepfake, it displays **"Fake"** otherwise **"Real"**.

Operating Environment

DEEP FAKE web application intends to detect deepfake videos using Flask for backend. We integrated our trained model in Fronted UI whic is build in ReactJS.

Figure 7.1: Tech stack of DEEP FAKE web application

Screenshots of Working Application

Figure 7.2: Home Page of DEEP FAKE application

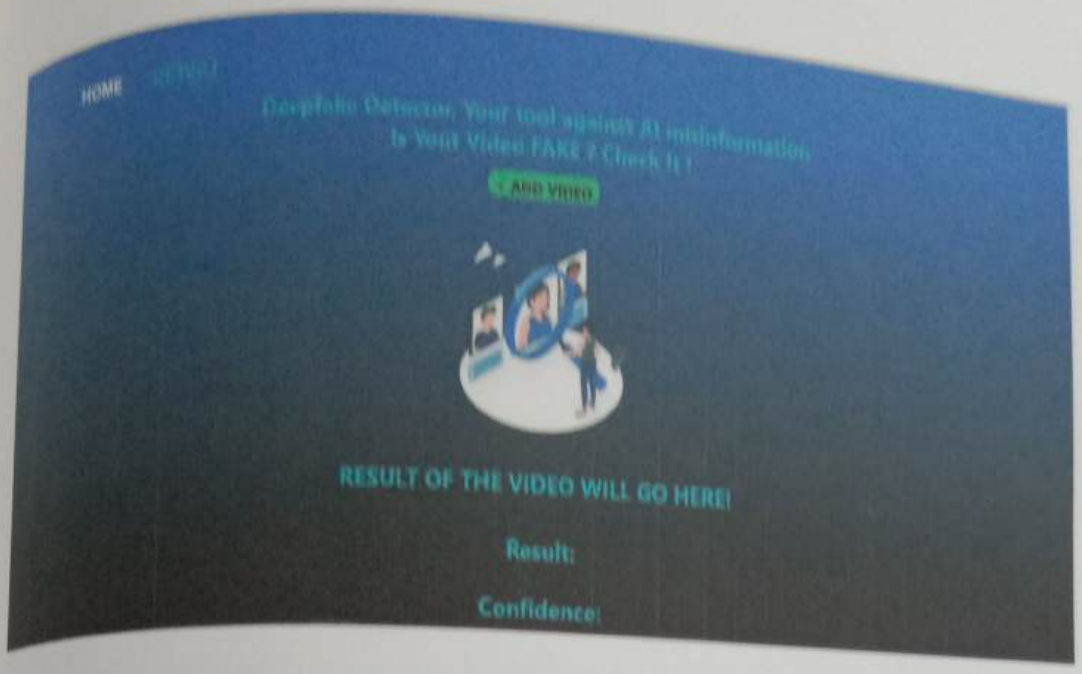


Figure 7.3: Detect Page of DEEP FAKE application

Figure 7.4: Uploading a video

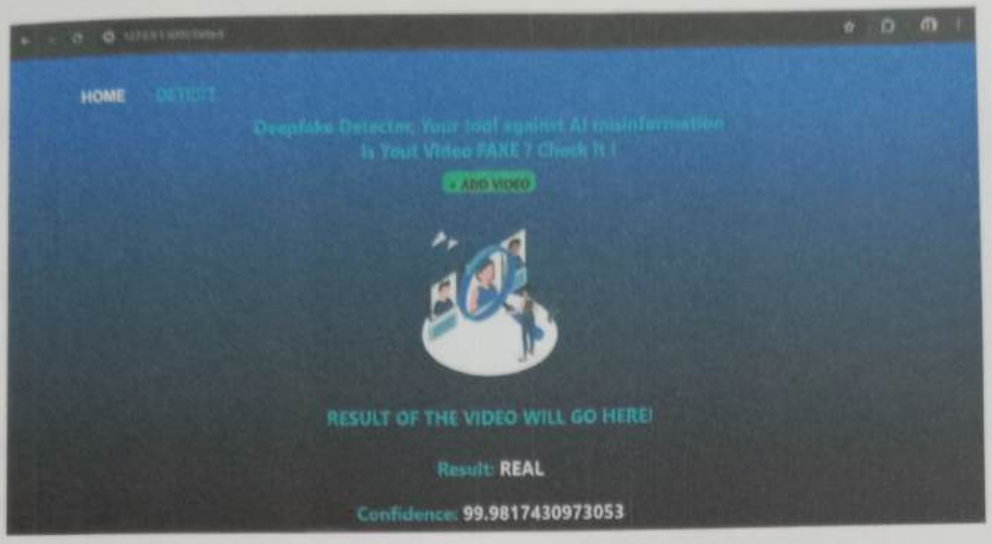


Figure 7.5: Result of input video which displays "REAL"

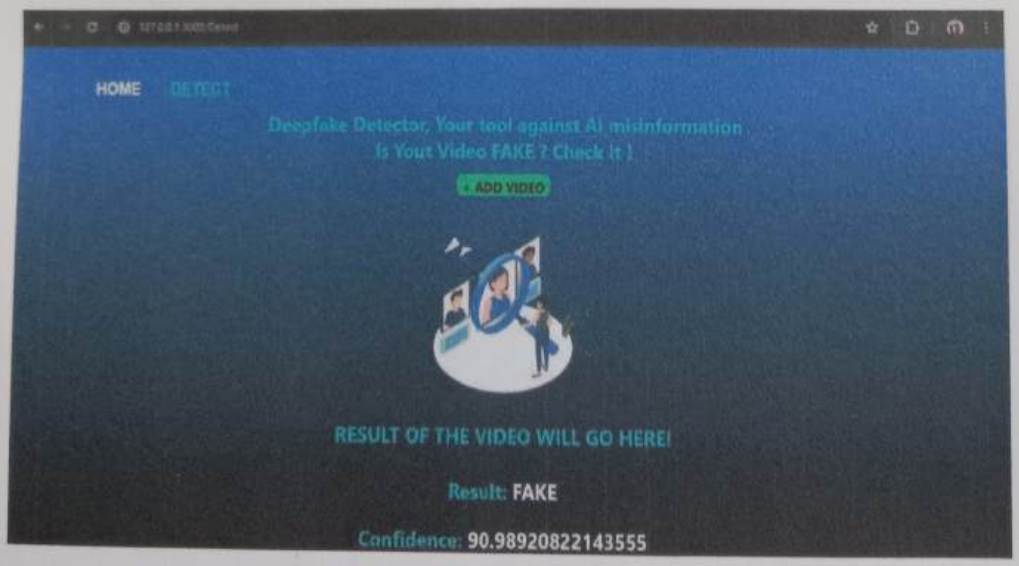


Figure 7.6: Result of another input video which displays "FAKE"

Chapter 8

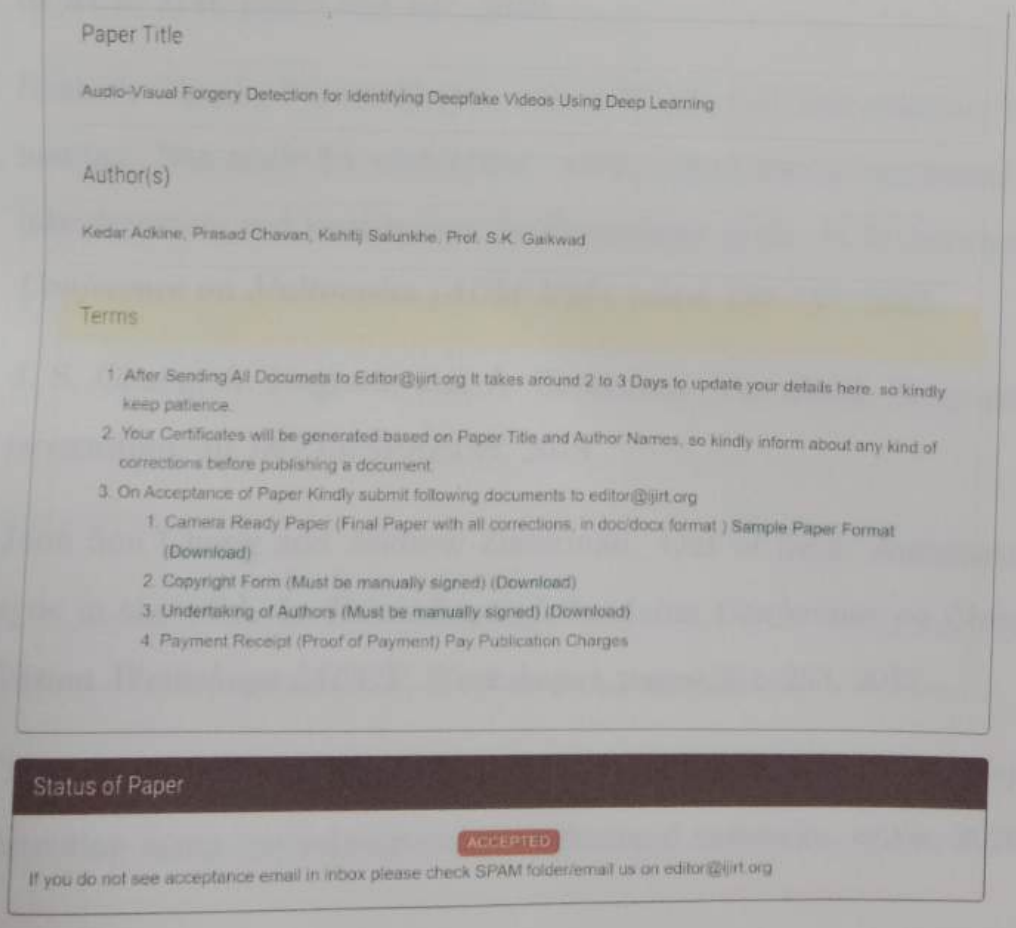
Conclusion

This study introduces a novel challenge associated with the generation and detection of content-driven deepfakes. Additionally, we propose a new approach for identifying temporal forgery within partially modified videos. Experimental results shows that our method surpasses the performance of existing cutting-edge techniques in similar domains.

Appendix A

Publication Details

Our paper titled- **Audio-Visual Forgery Detection for Identifying Deepfake Videos Using Deep Learning** has been accepted by the International Journal of Innovative Research in Technology (IJIRT) for publication.



Paper Title

Audio-Visual Forgery Detection for Identifying Deepfake Videos Using Deep Learning

Author(s)

Kedar Adkine, Prasad Chavan, Kshitij Salunkhe, Prof. S.K. Gaikwad

Terms

1. After Sending All Documets to Editor@ijirt.org It takes around 2 to 3 Days to update your details here, so kindly keep patience.
2. Your Certificates will be generated based on Paper Title and Author Names, so kindly inform about any kind of corrections before publishing a document.
3. On Acceptance of Paper Kindly submit following documents to editor@ijirt.org
    1. Camera Ready Paper (Final Paper with all corrections, in doc/docx format) Sample Paper Format (Download)
    2. Copyright Form (Must be manually signed) (Download)
    3. Undertaking of Authors (Must be manually signed) (Download)
    4. Payment Receipt (Proof of Payment) Pay Publication Charges

Status of Paper

ACCEPTED

If you do not see acceptance email in inbox please check SPAM folder/email us on editor@ijirt.org

Bibliography

[1] Aditya Bagchi, Jawadul Mahmood, Dilip Fernandes, and Ravi Kiran Sarvadevabhatla. Hear me out: Fusional approaches for audio augmented temporal action localization. *arXiv:2106.14118 [cs]*, 2021.

[2] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian. Not made for each other- audio-visual dissonance-based deepfake detection and localization. In *ACM MM*, pages 439–447, 2020.

[3] Kashish Chugh, Piyush Gupta, Aman Dhall, and Ramanathan Subramanian. Not made for each other - audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, pages 439–447, 2020.

[4] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

[5] Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Proceedings of the Asian Conference on Computer Vision Workshops (ACCV Workshops)*, pages 251–263, 2017.

[6] O. de Lima, S. Franklin, S. Basu, B. Karwoski, and A. George. Deepfake detection using spatiotemporal convolutional networks. *arXiv*, 2020.

[7] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv*, 2006.07397, 2020.

[8] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset, 2020.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[10] Y. He, B. Gan, S. Chen, Y. Zhou, G. Yin, L. Song, L. Sheng, J. Shao, and Z. Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In *CVPR*, pages 4360–4369, 2021.

[11] Y. Jia, Y. Zhang, R. J. Weiss, and et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *NeurIPS*, pages 4485–4495. Curran Associates Inc., 2018.

[12] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *CVPR*, pages 2889–2898, 2020.

[13] H. Khalid, S. Tariq, and S. S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv*, 2108.05080, 2021.

[14] I. Korshunova, W. Shi, J. Dambre, and L. Theis. Fast face-swap using convolutional neural networks. In *ICCV*, pages 3677–3685, 2017.

[15] T.-N. Le, H. H. Nguyen, J. Yamagishi, and I. Echizen. Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In *ICCV*, pages 10117–10127, 2021.

[16] Y. Li, X. Yang, P. Sun, et al. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *CVPR*, pages 3207–3216, 2020.

[17] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *ICCV*, pages 3889–3898, 2019.

[18] Tao Lin, Xueting Liu, Xiaoxiao Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3889–3898, 2019.

[19] Mohit Nawhal and Greg Mori. Activity graph transformer for temporal action localization. *arXiv:2101.08540 [cs]*, 2021.

[20] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019.

[21] C. Sanderson and B. C. Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *Advances in Biometrics*, pages 199–208. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[22] H. Su, W. Gan, W. Wu, Y. Qiao, and J. Yan. Bsn++: Complementary boundary regressor with scale-balanced relation modeling for temporal action proposal generation. *arXiv*, 2009.07641, 2021.

[23] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, and M. Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *ECCV 2020*, pages 716–731, 2020.

[24] K. Yang, P. Qiao, D. Li, S. Lv, and Y. Dou. Exploring temporal preservation networks for precise temporal action localization. *AAAI*, 32(1), 2018.

[25] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2914–2923, 2017.

[26] B. Zi, M. Chang, J. Chen, et al. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *ACM MM*, pages 2382–2390, 2020.