MMTFD: A multimodal detector for temporal forgeries detection

by

Yadvender Singh

May 10, 2024

A thesis submitted to the

Faculty of the Graduate School of

the University at Buffalo, The State University of New York

in partial fulfilment of the requirements for the

degree of

Master of Science

Department of Computer Science and Engineering

Acknowledgments

I express sincere gratitude to Dr. Nalini Ratha for his invaluable guidance and unwavering support during my academic endeavors. His knowledge and support have played a crucial role in shaping the outcome of this thesis, and I am really grateful for his guidance.

I express profound gratitude to my thesis committee member, Dr. Junsong Yuan, for his insightful ideas and valuable suggestions that will help advance this research forward.

I would also like to thank my friends and lab mates, Shawna, Sunil, and Bharat, for their continuous support and insightful contributions, which have been of great assistance in the completion of this work.

Most of all, I would like to convey my deep appreciation to my family for their constant support and unfailing confidence in me. Their unwavering support, comprehension, and forbearance have been the foundation of my scholarly endeavors, leading me over obstacles and motivating me to strive for distinction. I am profoundly appreciative of their affection and support, since they have served as the impetus for my achievements.

Table of Contents

List of Tables

List of Figures

Abstract

Deepfakes pose a significant challenge to the integrity of digital media, undermining trust in online content and raising doubts about the authenticity of visual information. Traditional detection methods typically analyse entire videos, often struggling to identify deepfake content when faced with temporal manipulations. The frame-by-frame detection techniques generally fail in precisely locating the temporal forged traces within a videos. Addressing this critical gap is imperative, prompting the exploration of advanced detection techniques capable of accurately pinpointing temporal alterations within videos. In our thesis, we introduce a 2 step innovative approach to deepfake detection, building upon recent advancements in audio and vision transformer architectures. Leveraging the powerful self-attention mechanisms inherent in these transformer models, first, we split a video in chunks of samples and the audio-visual features are extracted for the samples. Then these samples are classified using the Multi-Modal Temporal Forgery Detection (MMTFD) model and forged traces are identified that are randomly dispersed within videos. By utilizing an Audio transformer encoder and a Video transformer encoder, we meticulously process video segments, analyzing temporal and spatial inconsistencies across batched frames. This novel approach represents a significant leap forward in deepfake detection, as it effectively harnesses transformer models to enhance the reliability of multimedia content authentication systems, detecting Audio-Visual temporal forgeries with an accuracy of 96%. Our research contributes to the ongoing efforts to combat recent advancements in deepfakes by offering a robust and efficient method for identifying forged elements. By integrating cutting-edge technologies and methodologies, we strive to empower content authentication systems with the capability to detect Audio-Visual forgeries and localize temporal alterations accurately. Through this study, we aim to fortify the defenses against deepfake threats, ultimately preserving the integrity and trustworthiness of digital media in an era fraught with misinformation and manipulations.

Chapter 1

Introduction

In an age characterized by unparalleled progress in digital technologies, the ability to create exceptionally lifelike images and videos has skyrocketed, fueled by state-of-the-art computer graphics and artificial intelligence algorithms. Although this rise has numerous practical uses, it has also introduced a new domain of privacy and security issues. The primary issues revolve around the phenomenon known as deepfake, which combines the concepts of "deep learning" with "fake." Deepfake technology enables the seamless manipulation of photos or videos by superimposing or concealing one person's likeness with another's, modifying not only their appearances but also their voices and facial expressions. Utilizing advanced methods based on deep learning and artificial intelligence, the manipulation of deepfakes has become extremely challenging for humans to detect. This manuscript aims to clarify the notion of deepfake, examining its different forms and studying both the techniques used to create it and the methods employed to detect it.

Crafting a convincing deepfake, particularly for disseminating misinformation or fake news—such as a politician delivering a speech or issuing a statement—demands meticulous manipulation of both video and audio channels. Advancements in text-to-speech (TTS) and voice conversion (VC) algorithms have made it easier to create synthetic human speech. This indicates a future where audio will be as important as video in detecting deepfakes. This work focuses on exploring the complex connection between these two modalities, which is essential for detecting audio-visual deepfakes.

Previous efforts have predominantly centered on identifying visual anomalies and 'fingerprints' across various generative frameworks or pinpointing local texture inconsistencies

resulting from face swapping. Alternatively, some approaches rely on biometric signals, such as detecting unique facial motion patterns specific to individuals, though such identity-specific methods face limitations in generalization to new identities. In order to adopt a more comprehensive strategy, [1] utilizes the significant connection between the movements of the lips (viseme) and the enunciated syllables (phoneme) that are observed in human speech. This synchronization breaks out at subtle intersections when either modality is proven false. The discrepancies in lip movements and syllables occur because of distortions caused by face swapping or lip-syncing. Moreover, phonemes produced by text-to-speech (TTS) systems frequently lack distinct enunciation that may be synchronized with facial movements, which serves as a crucial indicator for identifying audio-visual deepfakes.

However, these audio-visual detection algorithms mostly focus on confirming the genuineness of complete videos, while neglecting the identification of manipulated fragments. Just a minor change to a few words can significantly change the interpretation of a statement. In an authentic video, the speaker says, "We must work together to ensure equality for all citizens." Alternatively, if the video and audio associated with the word "equality" were replaced with those representing "inequality," the sentence would express a completely contrasting impression. The orchestrated manipulations of this kind, especially when involving extreme utterances, have incalculable repercussions. This highlights a new requirement for detectors: not only to determine the authenticity of the video but also to accurately identify the specific time points when altered segments are present in the manipulated content.

Despite considerable efforts directed towards audio-visual temporal forgery localization, this remains a formidable challenge. Consequently, there's a pressing need for the development of methods capable of producing more precise and dependable forgery boundaries, a need that motivates our current work. Building upon insights gleaned from prior research endeavors, we introduce a novel MMTFD model for audio-visual temporal forgery detection in videos. Our contributions are delineated as follows:

- We present a comprehensive pipeline designed to preprocess and segment videos into

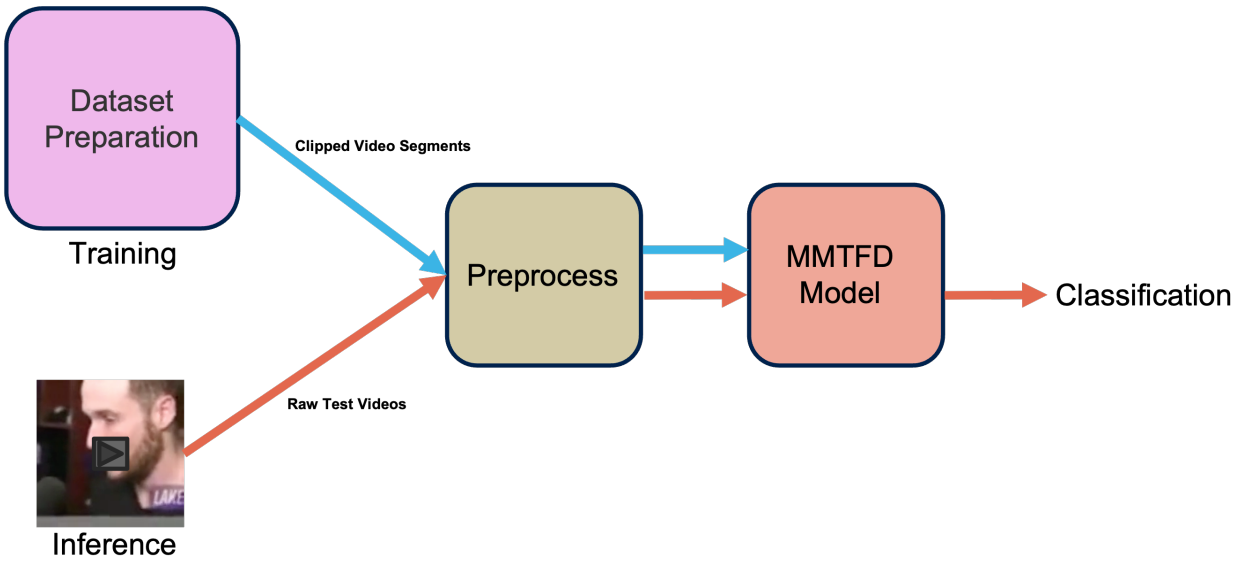


Figure 1.1: An overview of end-to-end training and inference pipeline.

smaller samples, facilitating the extraction of video and audio embeddings from these segments.

- We introduce an application of embedding fusion techniques to integrate audio and visual embeddings generated by transformer models: VideoMAE[2] and AST[3]. These fused embeddings are then utilized to classify samples into one of four categories of deepfake: Fake Audio Only, Fake Video Only, Both Audio and Video Fake, or Real.

Chapter 2

Related Work

Within the field of media synthesis, the phrase "Deepfake" refers to the application of deep learning methods to create deceptive media output, combining the concepts of "Deep Learning" and "Fake"[4]. The emergence of Deepfakes poses a significant difficulty because of the progress made in artificial intelligence, which allows for the production of media that blurs the distinction between what is real and what is artificially created. The notion of Deepfakes became widely known in 2017 when an unidentified Reddit user, using the alias "deepfakes," distributed altered media content showcasing celebrity faces that had been switched.

The widespread existence of hyper-realistic counterfeit material has made the task of manually detecting them extremely challenging. As a result, scientists in the field of machine learning have shifted their focus toward creating techniques to distinguish genuine media from altered versions. Several media forensics tools have been developed with the purpose of verifying different types of media, such as photos, videos, text, and audio, in order to detect instances of falsification or malicious intent. This study specifically concentrates on the identification of counterfeit videos, more frequently referred to as deepfake films.

2.1 Unimodal Deepfakes Detection

Unimodal Deepfake Detection refers to the process of identifying and distinguishing deepfake content using a single type of data or modality. In the past, methods for detecting forged videos have mostly used single-mode techniques, such as analyzing face features, scrutinizing images or frames, and identifying statistical abnormalities and visual anomalies for classifi-

cation. In Matern et al [5], the detection of forged videos relies on the identification of visual anomalies, such as differences in lighting, mismatched eye hues, and inconsistencies in the eye and tooth areas.

In addition, Bayar et al [6] and Afchar et al [7] suggested high-level and mesoscopic characteristics, respectively. Nguyen et al [8] introduced a capsule network, Chollet et al [9] developed an XceptionNet, and Zhou et al [10] designed a two-stream convolutional neural network (CNN) for detecting forgery. Although the focus of these efforts is mostly on visual analysis, the importance of audio data cannot be emphasized enough. Audio plays a crucial role, as automatic speaker verification (SV) systems are widely used to confirm speaker identities.

Nevertheless, the weaknesses of existing SV systems have been revealed, as they are susceptible to manipulation through audio signal alteration. Prior studies have extensively examined this matter, suggesting several solutions.

2.2 Multimodal Deepfakes Detection

Multimodal detection combines the strengths of many modalities to improve detection skills by taking advantage of their complementing characteristics. Empirical research confirms that combining information from both aural and visual sources leads to better performance compared to depending exclusively on data from a single modality.

Several approaches have been proposed for addressing the task of detecting deepfake content. Chugh et al. [11] have suggested a successful method that involves assessing the consistency of emotional characteristics obtained from both aural and visual sources. This approach employs emotional cues present in the information to determine its authenticity. Mittal et al. [12] devised a technique to evaluate the disparities between auditory and visual modalities for identifying probable instances of deepfake manipulation.

In addition, Zhou et al. [13] presented a sophisticated integrated audiovisual model that

utilizes the inherent synchronization between auditory and visual stimuli. This approach aims to determine the truthfulness of a certain movie by carefully analyzing the consistency between several modalities. This research demonstrates a sophisticated understanding of how audio and visual elements interact in videos, providing an effective strategy for precisely detecting deepfake content.

In addition, Cheng et al. [14] did a study that aimed to detect deepfakes by examining the relationship between facial and audio data through voice-face matching. LAV-DF [15] dataset introduced a technique for detecting alterations in audio-visual content by examining temporal boundaries. Furthermore, Agarwal et al. [1] detected artifacts by analyzing the temporal variations in mouth structure in relation to spoken phonemes. However, these attempts mostly focused on the explicit representation of information across distinct modalities, while neglecting the implicit integration of non-synergistic aspects. Furthermore, they often saw audio as additional signals for supervision, overlooking the possibility of audio forgeries, which commonly occur in real-life scenarios. The possible correlation between several senses in multimodal deepfake detection has not been extensively investigated or exploited.

2.3 Temporal Deepfakes Detection

Temporal deepfakes, which are a specific type of modified media, pose unique difficulties because to their temporal coherence and consistency. To detect temporal deepfakes, it is necessary to use approaches that take into consideration the dynamic changes in information over time, including both visual and auditory clues.

Forged videos often exhibit anomalies in genuine physiological traits, leading to disparities with actual humans. To address this problem, researchers focus on assessing the reliability of the physiological features of artificially generated faces displayed in films. Li et al. [16] suggest using blinking patterns and blink frequency as measures to assess the validity of a video. Yang et al. [17] examine disparities in head poses by comparing the variations in

head poses predicted using all facial features with only the markers in the central area.

Inter-frame inconsistency detection techniques are specifically designed to identify disparities between images in successive frames or frames with defined time intervals. Gu et al. [18] emphasize the significance of inter-frame image discrepancies by the intensive capture of neighboring frames. Yin et al. [19] employ a Dynamic Fine-grained Difference Capturing module and a Multi-Scale Spatio-Temporal Aggregation module to accurately depict spatio-temporal inconsistencies. Yang et al. [20] address the issue of identifying deepfakes by approaching it as a problem of graph classification. Their primary emphasis is to analyze the interconnections among various facial regions in successive frames. Furthermore, Choi et al. [21] employ differences in style variables across frames to develop a style attention module capable of detecting inconsistencies in style latent variables.

Multimodal detection algorithms utilize data from various variables to create conclusions, surpassing the distinctions found in individual images or audio. These strategies give priority to the transmission of previous information from both visual and aural modes. POI-Forensics [22] uses contrastive learning to verify the authenticity of audio-visual content, while AVoiD-DF [23] combines spatiotemporal information to merge multimodal features. Agarwal et al. [24] suggest a forensic method that uses both static and dynamic auditory ear features to identify counterfeit faces. Research in forgery detection is currently focused on multimodal detection systems, which offer advanced capabilities for spotting deepfake manipulation.

Chapter 3

Datasets

3.1 Introduction

The process of selecting datasets plays a crucial role in the development and evaluation of algorithms specifically geared to identify deepfake movies. This section provides a summary of the progression of deepfake datasets throughout time and an overview of datasets used in this research. The deepfake datasets are categorized into four generations based on the characteristics and advancements in forgery techniques that they embody.

3.1.1 First Generation Datasets

The initial generation includes datasets such as DF-TIMIT [25], UADFV [17], SwapMe, and FaceSwap [47]. DF-TIMIT carefully selects and organizes 16 sets of persons who have similar physical appearances from the VidTIMIT database. This process results in a collection of 640 movies where the faces of the participants have been exchanged. The UADFV dataset has a total of 98 videos, with 49 being authentic and 49 being artificially created using the FakeAPP software. SwapMe and FaceSwap utilize two face-swapping software apps to generate counterfeit photos from a set of 1005 authentic images taken in 2010.

3.1.2 Second Generation Datasets

The second generation datasets demonstrate enhancements in both size and quality when compared to their predecessors. The Google DeepFake Detection[26] dataset consists of 3,068

counterfeit videos produced using five alteration techniques that are publicly accessible. The Celeb-DF[27] dataset comprises 590 authentic YouTube videos showcasing famous individuals, together with 5,639 altered video snippets. FaceForensics++ [28] comprises a collection of 4000 counterfeit videos that have been altered using four different techniques (DeepFakes, Face2Face, FaceSwap, and NeuralTextures), in addition to 1000 authentic YouTube videos.

3.1.3 Third Generation Datasets

The latest datasets for face forgeries are considered the third generation, distinguished by their large size and variety. DeeperForensics-1.0 [29] consists of a collection of 60,000 videos specifically designed for detecting instances of face forgery in real-world scenarios. DFDC[30] has a collection of more than 100,000 video clips obtained from 960 actors who were paid. These clips were created using several methods of replacing faces in the videos. DFFD [31] introduces spatial forgery annotations, but only for binary masks without manipulation density.

For practical purposes, apart from classification tasks, it is essential to be able to identify and locate altered sections or segments inside images or movies. Although certain datasets, like DFFD, tackle these tasks, additional progress is required to offer extensive annotations for the identification of manipulation in real-world situations.

3.1.4 Fourth Generation Datasets

Fourth generation datasets provide notable progress in the identification of deepfake videos, especially in terms of accurately pinpointing the timing and location of changes within multimedia content. Contrary to previous datasets, these new datasets are developed using LLMs such as ChatGPT and are designed to capture subtle alterations that are hidden inside genuine content segments, enabling more precise detection.

In the past, alterations were primarily restricted to the visual mode. Nevertheless, as advancements occurred, audio manipulations and audio-visual manipulations were employed

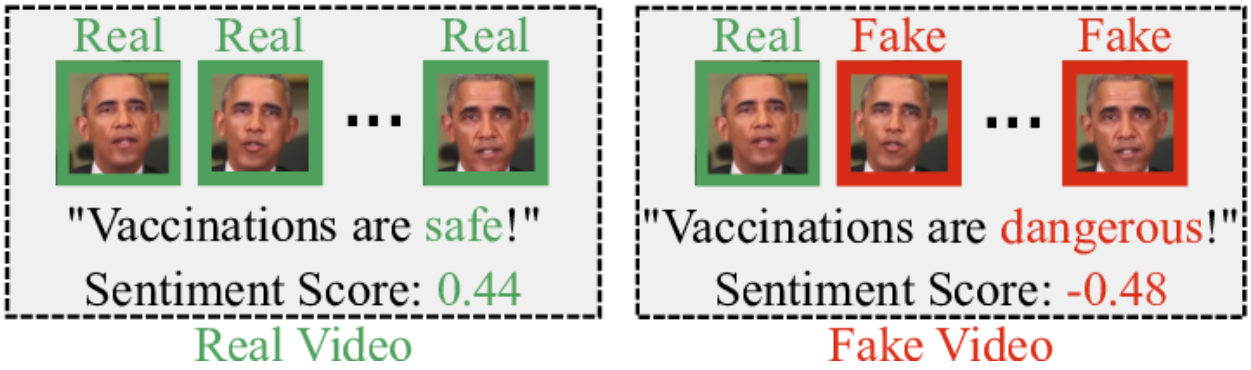


Figure 3.1: A sample of LAV-DF[15] dataset showcasing how temporal manipulation can change the sentiment of a video.

to enhance the intricacy of activities. In 2022, a notable achievement in the development of deepfake datasets that focus on temporal localization was reached with the introduction of LAV-DF[15]. The objective of LAV-DF was to pinpoint modified segments within multimedia data, to establish the foundation for more advanced detection methods. The AV-Deepfake1M [32] dataset is an advanced collection of data that represents the highest achievement in fourth generation datasets. This dataset overcomes the constraints of earlier datasets by greatly improving the quality, diversity, and scale of information specifically designed to detect temporal deepfakes.

3.2 Datasets Used

3.2.1 LAV-DF

Localized Audio Visual DeepFake (LAV-DF) is a significant compilation of audio-visual deepfake data. The fundamental idea behind deepfake generation in LAV-DF is based on the theory that modifying important words in a transcript might have a significant impact on how it is perceived. More precisely, this manipulation seeks to alter the emotional tone of the transcript by replacing carefully chosen words with their opposite meanings. This substitution tactic results in a significant alteration in the emotional tone of the sentence. Following are the steps for generating deepfakes in this dataset:

Data Sourcing: The real videos are initially collected from the Vox-Celeb2 [33] dataset, which is a large library of facial videos. The films are subjected to facial tracking and cropping at a resolution of 224×224 pixels using the facial detector outlined in [55]. The selection is based on confidence scores acquired from the Google Speech-to-Text service. The transcripts for manipulation are produced with a similar service.

Transcript Manipulation: After obtaining the authentic movies, the transcript of each video is examined to find tokens that will be substituted which when replaced, will have the greatest impact on the overall attitude.

Audio Generation: Afterwards, the speaker's style is used to generate the relevant audio. The preferred method for audio synthesis is SV2TTS [34].

Video Generation: The artificially created false audio is used as input to generate equivalent fabricated video frames. The work at hand utilizes Wav2Lip [35], a tool specifically designed for facial reconstruction.

The LAV-DF dataset consists of three different types of manipulated data:

1. Fake audio and fake video
2. Fake audio and real video
3. Real audio and fake video

This dataset comprises 136,304 videos encompassing a diverse range of content. Out of the total number of videos, precisely 36,431 are real, while 99,873 consist of segments that have been manipulated to produce deepfake videos. The collection has 153 unique identities, contributing to its depth and diversity.

Data Split

Training set consists of 78,703 videos that represent 91 different identities.

Validation set consists of 31,501 videos, each belonging to 31 unique identities.

Testing Set comprises of 26,100 films that encompass 31 unique identities.

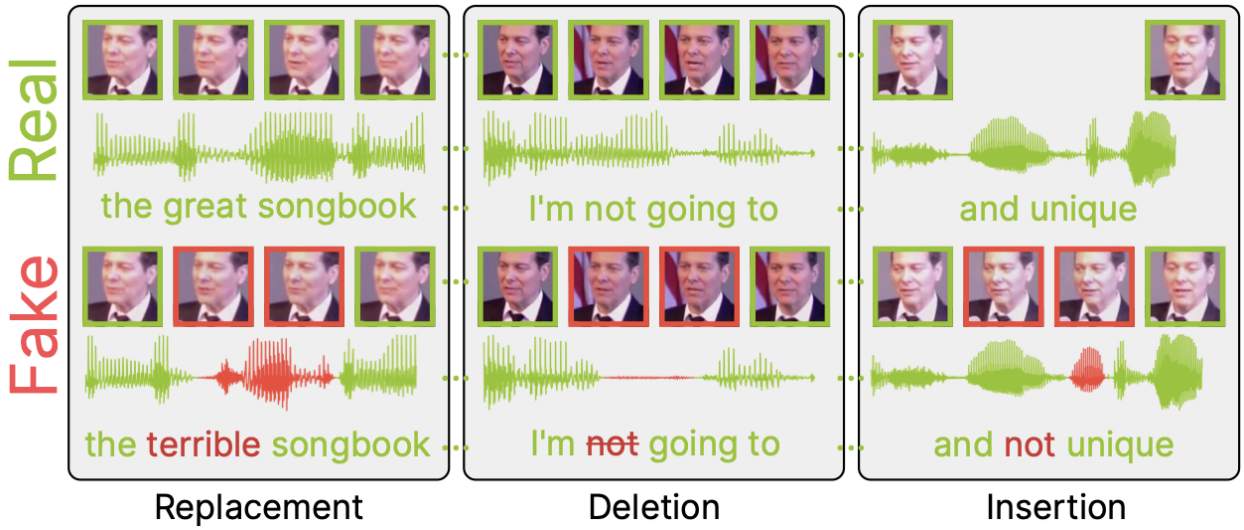


Figure 3.2: A sample of AV-Deepfake1M[32] dataset showcasing different type of temporal forgeries.

3.2.2 AV-Deepfake1M

Dataset	SECS	SNR	FAD
LAV-DF	0.984	7.83	0.306
AV-Deepfake1M	0.991	9.39	0.088

Table 3.1: Audio Quality Comparison[32]

The AV-Deepfake1M dataset is a noteworthy addition to the collection of audio-visual deepfake datasets. This dataset is highly thorough due to its large volume, diverse content, and careful curation, which places it at the forefront of audio-visual benchmarking initiatives. AV-Deepfake1M, similar to LAV-DF, employs a meticulously designed three-step process to create deepfakes that are focused on content.

Dataset	PNSR	SSIM	FID
LAV-DF	33.06	0.898	1.92
AV-Deepfake1M	39.49	0.977	0.49

Table 3.2: Video Quality Comparison[32]

Transcript manipulation: The initial stage involves making alterations to the transcripts of genuine videos to include adjustments that are driven by the content. To achieve this,

the original transcripts are altered using ChatGPT, leveraging its natural language processing capabilities to alter transcripts in a contextually appropriate way, allowing for realistic alterations to the material.

Audio Manipulation: Afterwards, superior audio is generated to replicate the original speaker's manner of speaking. The VITS approach is used to improve the quality and coherence of audio for a specific group of subjects. To incorporate a range of different audio styles in the dataset, the YourTTS text-to-speech approach is used for the remaining subjects, regardless of their identification. This dual approach ensures the incorporation of both superior quality and varied audio content, hence augmenting the overall genuineness of the deepfake videos.

Video Manipulation: The final stage of the generating pipeline is specifically focused on the creation of visual content. TalkLip is a specialized tool exclusively developed for producing deepfake movies that effectively synchronize lip movements without requiring any training data. TalkLip's features ensure that the movies created have precise lip synchronization and facial expressions, ensuring consistency between the adjusted audio and visual parts. This dataset consists of 3 different types of manipulated data similar to LAV-DF.

Data Split

Training set consists of 186,666 real videos and 559,514 fake videos of 1657 subjects.

Validation set consists of 14,235 real videos and 43,105 fake videos with 1657 subjects.

Testing Set comprises of 85,820 real videos and 257,420 fake videos of 411 subjects.

Chapter 4

Methodology

In this chapter, we will discuss the end-to-end training and inference pipeline from data preparation to train the novel model to classify multimodal temporal forgeries.

4.1 Data Preparation

In the context of temporal forgeries, where the occurrence of manipulated segments within videos is irregular, a preprocessing step is imperative to address the disproportionate distribution between real and manipulated segments. Direct utilization of entire videos could lead to an imbalanced dataset skewed towards real video segments. Therefore, a data preparation step is undertaken to ensure dataset balance.

To achieve this, videos from the dataset are clipped near the identified fake segments. Specifically, for training and validation datasets, segments are clipped one second before and one second after the fake segment, effectively balancing the dataset.

The dataset is prepared utilizing metadata files associated with the respective datasets. The attribute denoting fake segments or periods is utilized to delineate and extract clips, which are subsequently used for training and validation purposes.

4.2 Data Preprocessing

The videos are initially loaded and sliced into smaller chunks comprising 16 frame samples each, alongside their respective audio samples. After loading the samples, the visual and

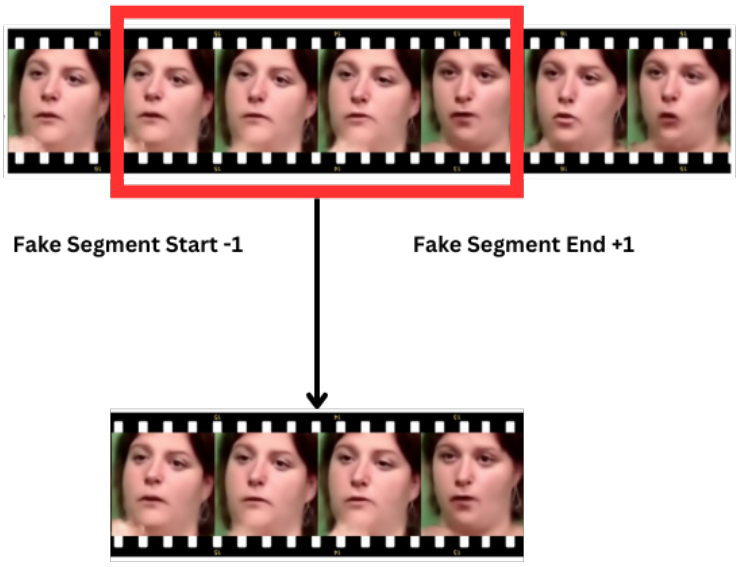


Figure 4.1: Data Preparation: Training and Validation videos are sliced based on fake segments available in dataset metadata.

audio data is preprocessed for each subclip of the video.

4.2.1 Video Preprocessing:

The following transforms are applied to the video data of each subclip to enhance model performance and generalization:

1. Resize to (224x224): Videos are resized to a standard dimension to ensure uniformity across the dataset.

2. Random cropping: Random cropping is performed to extract diverse spatial features from the videos, augmenting the dataset.

3. Uniform Temporal Subsampling: Uniform temporal subsampling ensures that temporal information within the video segments is adequately represented.

4. Normalize: Video data is normalized to facilitate model convergence and mitigate training instability.

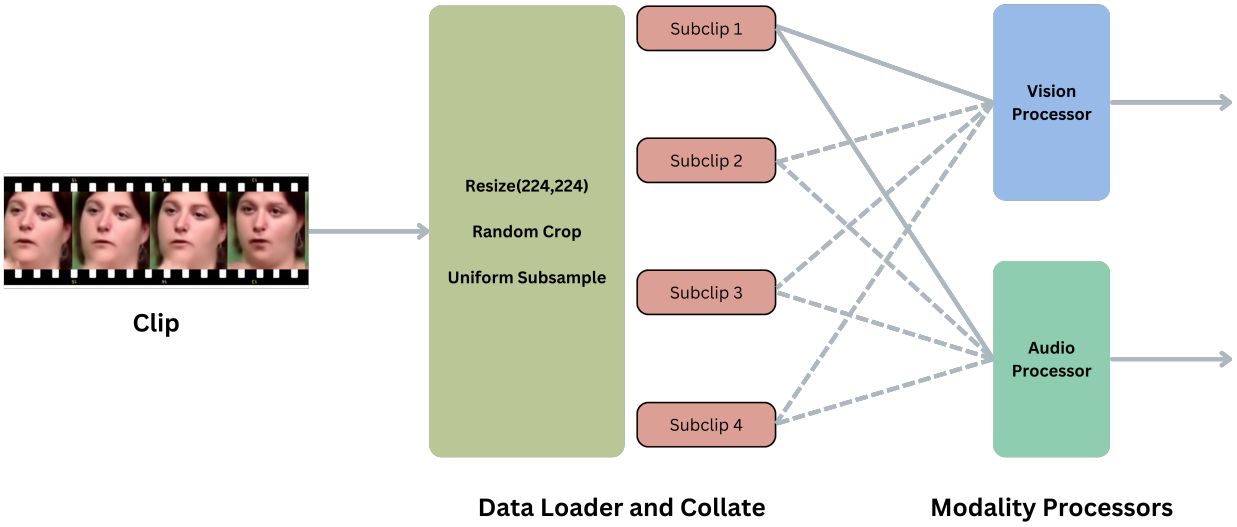


Figure 4.2: Data Preprocessing: Videos are split into small chunks of 16 frames and respective audio frames which are passed through audio and video processors.

5. Random Horizontal Flip: Horizontal flipping is applied randomly to augment the dataset and enhance model robustness against horizontal variations.

4.2.2 Audio Preprocessing:

1. Audio data undergoes preprocessing using the AST Feature Extractor.

2. This feature extraction process extracts mel-filter bank features from raw speech and pads/truncates them to a fixed length and normalizes them using a mean and standard deviation ensuring essential audio features are captured and represented effectively for subsequent model training.

4.3 Model

With the advancements in audio-visual deepfake generation as discussed previously, the fusion of audio and visual information has emerged as a potent approach for enhancing the performance of deepfake classification tasks. Multimodal models capable of processing both audio and video inputs have garnered considerable attention for their ability to extract rich,

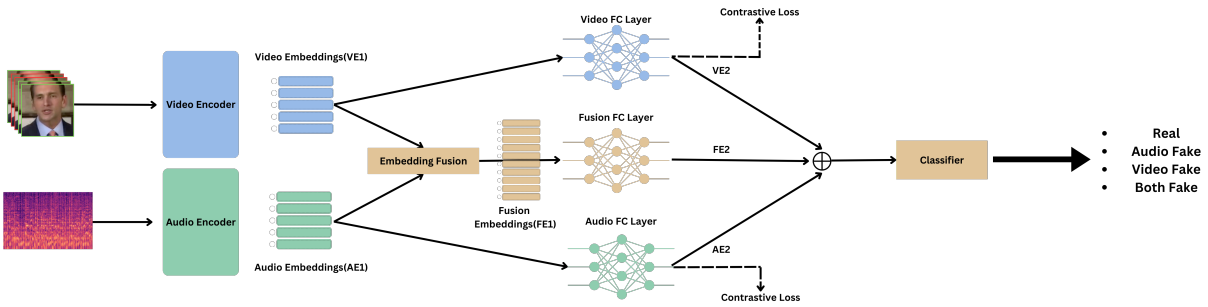


Figure 4.3: MMTFD: Novel Multimodal forgery detector.

complementary features from heterogeneous data sources. In this research, we are introducing a novel Multimodal Temporal Forgery Detector(MMTFD) model, which can process chunks of audio and visual informations to extract meaningful information and classify the video samples as fake or real based on the two modalities. Hence it can detect following four classes of temporal traces of forgeries in videos:

1.) Fake video-Fake audio 2.) Fake video-Real audio 3.) Real video-Fake audio 4.) Real video-Real audio

Once the audio and vision processors have completed preprocessing, the resulting data is then fed into the model, as shown in Figure 4.3. The video frames are processed by a video encoder, while the audio spectrogram is processed by an audio encoder. The video embeddings (VE1) and audio embeddings (AE1) are analyzed separately and then fused together (FE1) for more informative representation learning and to elucidate semantic connections between the audio and video in the subclip and . These embeddings are then fed into specialized fully connected layers to extract relevant information. The video features (VE2) and audio features (AE2) are produced using the Video FC Layer and Audio FC Layer, respectively. These attributes are subsequently employed to calculate pairwise contrastive loss. Ultimately, the feature embeddings AE2, VE2, and FE2 are combined and passed through a classifier to classify temporal forgeries traces into one of the four categories.

Video Encoder: We employ Video Masked Autoencoders (VideoMAE) as our video encoder due to their exemplary performance across various video classification benchmarks.

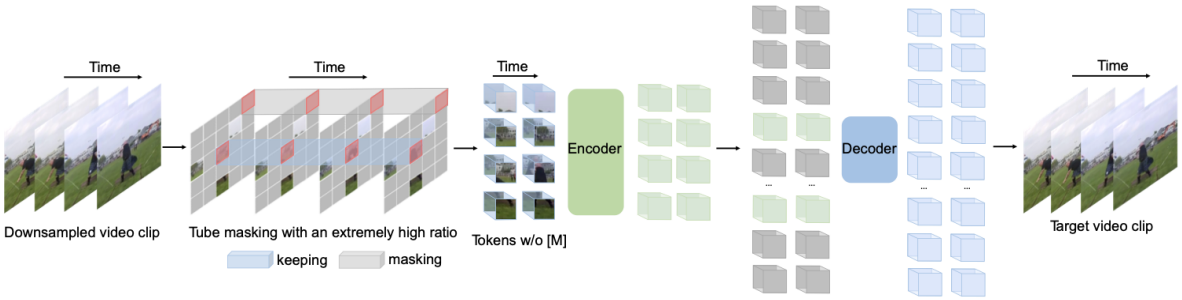


Figure 4.4: VideoMAE[2] model used as Video Encoder.

VideoMAEs serve as data-efficient learners, necessitating less data for effective training. Leveraging a customized tube masking design with an exceptionally high ratio enables meaningful self-supervised tasks, enhancing the ability of learned representations to capture useful spatiotemporal structures. The architecture of VideoMAE is illustrated in Figure 4.4.

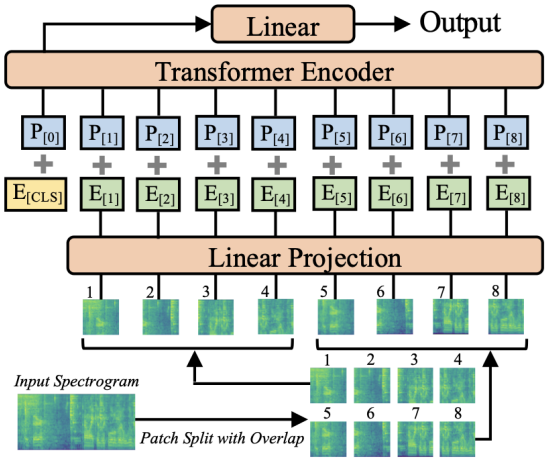


Figure 4.5: AST[3] used as Audio Encoder

Audio Encoder: For audio encoding, we adopt the Audio Spectrogram Transformer (AST), which has demonstrated state-of-the-art results in audio classification tasks. The architecture of AST is similar to Vision transformer which operates on a 2D audio spectrogram. This spectrogram is partitioned into a sequence of 16×16 patches with overlap, subsequently linearly projected into a sequence of 1-D patch embeddings. Each patch embedding is augmented with a learnable positional embedding, with an additional classification

token prepended to the sequence. The resultant output embedding serves as input to a Transformer, and the output of the classification token is utilized for classification via a linear layer. The architectural layout of AST is depicted in Figure 4.5

Contrastive Loss: Contrastive loss functions as a margin-based loss function, substituting Cross Entropy Loss in classification tasks. This loss function operates by pulling together clusters of points belonging to the same class in embedding space, while simultaneously pushing apart clusters of samples from different classes. Utilizing the output of paired visual and audio classifiers, contrastive loss computation yielded an average 1% enhancement in accuracy and precision across both intra-dataset and inter-dataset evaluations.

Chapter 5

Experiments and Results

5.1 Experiments

During the development of our model, we performed a sequence of experiments to enhance the fusion techniques used to combine audio and visual embeddings. We conducted investigations into several approaches for combining these embeddings, as well as exploring other loss functions to improve the model's performance.

5.1.1 Embedding Fusion Techniques

During the development of the model, various embedding fusion techniques were explored to integrate audio and visual information effectively. These techniques included concatenation, merging, and channel-level fusion.

1. Merging: Initially, we attempted to merge embeddings by performing element-wise summation. However, this approach did not yield satisfactory results.

2. Channel-Level Fusion: Here, we fused audio and visual embeddings to generate 2-channel embeddings, and observed promising outcomes in subsequent evaluations.

3. Concatenation: Concatenating both embeddings into a single channel emerged as the most successful fusion technique in our experiments, demonstrating superior performance over other methods.

5.1.2 Embedding Processing

Attention Module vs. Fully Connected Layer: Comparing these two techniques, we found that employing a fully connected layer yielded superior results in subsequent evaluations.

5.1.3 Loss Functions

To optimize our model's performance, we evaluated different loss functions, specifically focusing on Cross Entropy and Contrastive loss functions.

1. Cross Entropy Function: As a standard choice in classification tasks, the Cross Entropy function was thoroughly evaluated to establish a baseline performance for comparison with other loss functions.

2. Contrastive Loss Function: Utilizing pairwise inputs from embeddings generated by the video and audio fully connected layers, we observed an average improvement of 1% in classification accuracy compared to the Cross Entropy function.

5.2 Training Pipeline

During the training phase, we utilized a single Nvidia A100 80 GB GPU to optimize the parameters of our model for audio-visual classification tasks. Our training datasets consisted of subsets extracted from two primary sources: AV-Deepfake1M and LAV-DF.

5.2.1 Hardware Setup:

Our training infrastructure relied on a single Nvidia A100 80 GB GPU, known for its high computational power and efficiency in handling complex deep learning tasks. This GPU provided the necessary computational resources to train our model efficiently while ensuring rapid iteration and experimentation.

5.2.2 Datasets:

We curated subsets from two prominent datasets to train our model effectively:

AV-Deepfake1M: This dataset comprises a vast collection of videos encompassing various forms of deepfake content. We extracted a subset consisting of approximately 20,000 videos to train our model on tasks related to detecting manipulated audio-visual content.

LAV-DF: The LAV-DF dataset contains videos with manipulated audio-visual content, offering a diverse range of scenarios and manipulation techniques. We utilized a subset of approximately 30,000 videos from this dataset to further enrich our training data and enhance the model's ability to generalize across different forms of manipulation.

5.2.3 Training Procedure:

With our hardware setup and dataset composition in place, we commenced the training process. The model underwent training for a total of 20 epochs on each dataset, allowing it to iteratively learn from the training samples and adjust its parameters to improve performance.

5.2.4 Checkpoint Saving:

Throughout the training process, we implemented a checkpoint mechanism to monitor the model's performance and save the best-performing checkpoints based on accuracy. This approach ensured that we retained snapshots of the model's state at various stages of training, enabling us to revert to the most optimal configuration if necessary and track its progression over time. Following are some training metrics:

5.3 Inference Pipeline

During the inference phase, our model undergoes a rigorous pipeline designed to ensure accurate and efficient analysis of raw test videos. The pipeline, shown in Figure 5.3 comprises

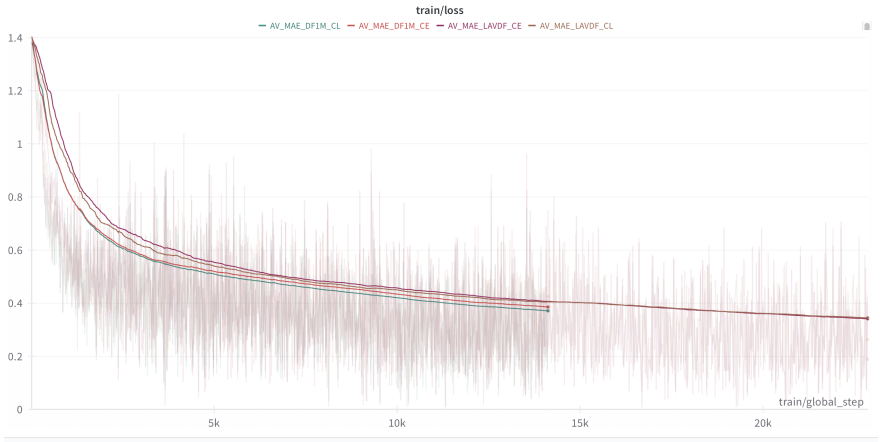


Figure 5.1: Training loss on Deepfake1M and LAV-DF using Contrastive Loss and Cross Entropy Loss

several key steps, each tailored to optimize performance and enhance the model's ability to generalize across diverse datasets.

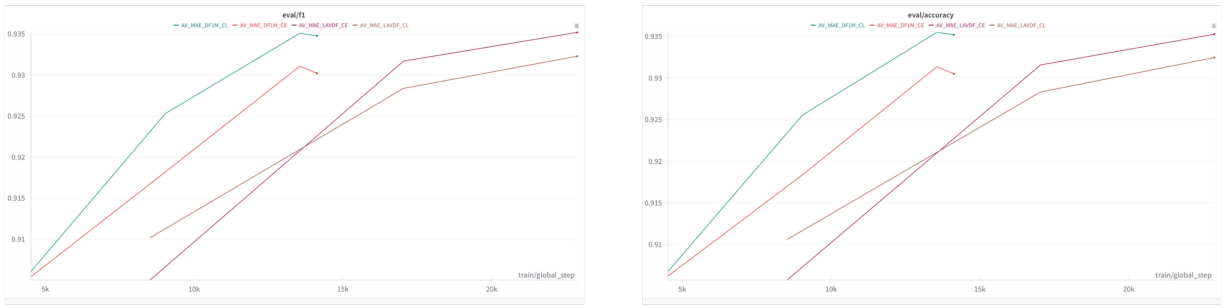


Figure 5.2: Training Evaluation Metrics(f1 and accuracy) on Deepfake1M and LAV-DF using Contrastive Loss and Cross Entropy Loss

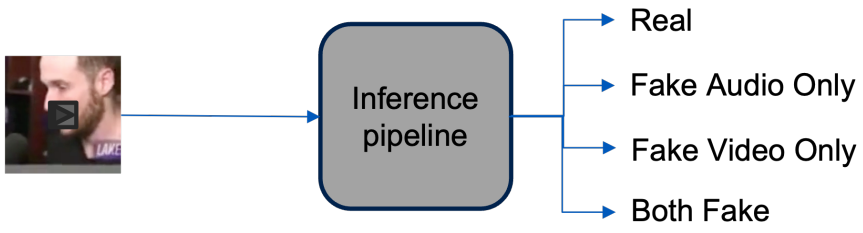


Figure 5.3: Raw videos fed into Inference Pipeline

5.3.1 Preprocessing:

Raw test videos are initially subjected to preprocessing steps to prepare them for analysis. One crucial aspect of preprocessing involves segmenting the videos into uniform subclips. This segmentation ensures that each segment contains consistent temporal information, facilitating more granular analysis by the model.

5.3.2 Subclip Segmentation:

The segmented subclips are created to maintain temporal consistency and standardize the input format for the subsequent stages of the pipeline. This step plays a pivotal role in ensuring that the model can effectively capture relevant features and patterns within each segment.

5.3.3 Model Inference:

Once the preprocessing is complete, the subclips are fed into the trained model for inference. The model utilizes its learned parameters and architectures to analyze each subclip independently and make predictions regarding the presence of specific attributes or classes within the video.

5.3.4 Timestamp Calculation and Class Assignment:

Following the inference stage, the model assigns timestamps to each subclip based on its index within the original video. These timestamps provide temporal context and facilitate post-analysis interpretation. Additionally, the model assigns an inferred class label to each subclip based on the predictions made during the inference process.

Chapter 6

Results

In this section, we present the comprehensive evaluation of our model using a range of performance metrics to assess its effectiveness in audio-visual classification tasks. We employ a multi-faceted approach, utilizing metrics such as:

1. Accuracy 2. Confusion Matrix 3. F1-score 4. Average Precision
5. Average Recall

These metrics provide a holistic understanding of the model's capabilities, offering insights into its classification accuracy, ability to handle class imbalances, and precision-recall trade-offs. Through rigorous evaluation, we aim to demonstrate the robustness and effectiveness of our proposed methodology in capturing the intricacies of audio-visual data for classification purposes.

6.1 Model Results and Analysis

Table 6.1 shows the results of evaluation on the same dataset while using different loss functions. It can be observed that the contrastive loss function helped improve model performance. Table 6.2 depicts the evaluation results on both the same dataset and cross-dataset

Dataset	Loss function	Accuracy	F1 Score	Precision	Recall
LAV-DF	Cross Entropy	0.9573	0.9572	0.9579	0.9573
	Contrastive Loss	**0.9693**	**0.9696**	**0.9704**	**0.9693**
AV-Deepfake1M	Cross Entropy	0.9207	0.9204	0.9207	0.9207
	Contrastive Loss	**0.9329**	**0.9326**	**0.9328**	**0.9329**

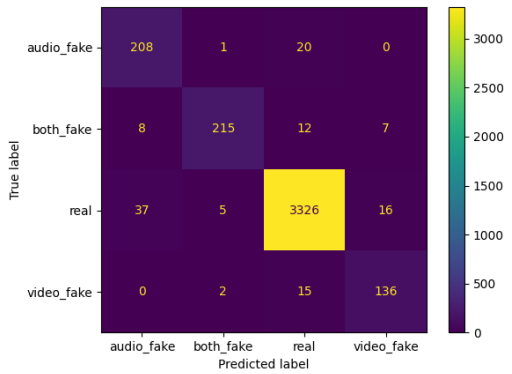
Table 6.1: Comparison of Model performance with Contrastive Loss and Cross Entropy Loss

scenarios. Notably, the model trained on LAV-DF exhibited superior performance during testing on the same dataset, yet faced challenges when tested on the cross-dataset Deepfake1M. Conversely, the model trained on AV-Deepfake1M demonstrated consistent performance across both same and cross-dataset evaluations, suggesting its adaptability to diverse manipulation techniques encountered in real-world scenarios. In Table 6.3, the confusion

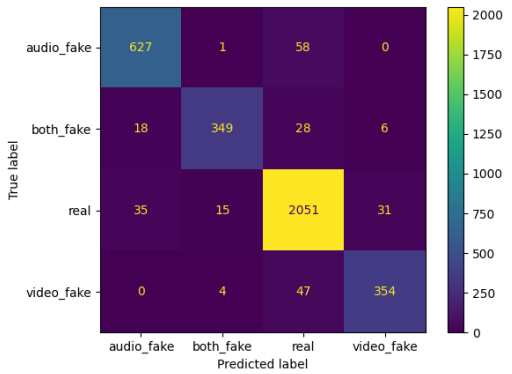
Training Dataset	Test Dataset	Accuracy	F1 Score	Precision	Recall
LAV-DF	LAV-DF	0.9693	0.9696	0.9704	0.9693
	AV-Deepfake1M	0.6667	0.5889	0.7283	0.6667
AV-Deepfake1M	LAV-DF	0.8760	0.8672	0.8619	0.8760
	AV-Deepfake1M	0.9329	0.9326	0.9328	0.9329

Table 6.2: Model Performance Metrics across same dataset and cross dataset

matrices show class-wise inference results. It can be observed here that in raw videos, the number of real samples is far higher than other class samples, and total accuracy



LAV-DF



AV-Deepfake1M

Table 6.3: Comparison of temporal forgery localization results against SOTA on the subset of LAV-DF dataset.

6.2 SOTA Comparison

In our evaluation against state-of-the-art (SOTA)(refer Table 6.4) methods on a subset of the LAV-DF dataset, several models were compared based on their performance in temporal forgery localization. AVFusion, a competitive baseline, achieved an average precision (AP)

of 0.6201 and an average recall (AR) of 0.6198. ActionFormer demonstrated enhanced outcomes with an Average Precision (AP) of 0.7948 and an Average Recall (AR) of 0.7038. Similarly, BA-TFD showed notable improvement compared to AVFusion, with an AP of 0.7690 and an AR of 0.6734. Expanding on this, BA-TFD+ significantly improved the outcomes, attaining an Average Precision (AP) of 0.9682 and an Average Recall (AR) of 0.8174. Our suggested model, MMTFD, achieved superior performance compared to other existing methods, with an Average Precision (AP) of 0.9704 and an Average Recall (AR) of 0.9693. These results demonstrate the efficacy of our technology in precisely identifying temporal forgeries, surpassing the current leading benchmarks in the field.

Model	**AP**	**AR**
AVFusion [36]	0.6201	0.6198
ActionFormer [37]	0.7948	0.7038
BA-TFD [15]	0.7690	0.6734
BA-TFD+ [38]	0.9682	0.8174
MMTFD(Ours)	**0.9704**	**0.9693**

Table 6.4: Comparison of temporal forgery localization results against SOTA on the subset of LAV-DF dataset.

Chapter 7

Conclusion and Future Work

In conclusion, our investigation into the realm of multimodal deepfakes and temporal forgery highlights the critical importance of developing robust detection mechanisms to combat their proliferation. We have elucidated the diverse landscape of forgery types, emphasizing the urgent necessity for advanced techniques capable of discerning increasingly sophisticated manipulations.

Our research underscores the efficacy of attention mechanisms inherent in Transformer models, particularly in capturing intricate spatio-temporal features present in modern deepfakes. Building upon this insight, we introduced the Multi-Modal Temporal Forgery Detection (MMTFD) model, specifically tailored to identify temporal audio-visual forgeries in digital media. By harnessing the power of Transformers, MMTFD represents a significant advancement in the field of forgery detection, offering enhanced accuracy and reliability in distinguishing manipulated content from authentic sources.

In the future, there are various possibilities for further investigation and improvement:

1. **Adversarial Robustness:** Investigate methods to enhance the adversarial robustness of the MMTFD model against sophisticated attacks designed to evade detection. Adversarial training, robust optimization techniques, and adversarial data augmentation could be explored to fortify the model's resilience to adversarial perturbations.

2. **Explore other Multimodal Tasks:**While our model demonstrates promising results

in temporal forgery detection, there exists a broader spectrum of multimodal classification tasks that could benefit from its capabilities. Future research could explore deploying the MMTFD model in diverse domains such as sentiment analysis, scene understanding, and event recognition.

3. **Optimizations for Resource Utilization:**To enhance scalability and deployment feasibility, exploring optimizations such as LoRA (Low-Rank Adaptation) and quantization techniques could be instrumental. These approaches aim to reduce resource utilization for both training and inference without compromising model performance, thereby making the detection framework more accessible and cost-effective in real-world applications.

In summary, our research not only contributes to the ongoing efforts in combating forgeries in digital media but also sets the stage for further advancements in multimodal detection methodologies. By continuing to innovate and adapt our techniques to emerging challenges, we can bolster our defenses against the proliferation of deceptive content in the digital landscape.

Bibliography

[1] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. "Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2020, pp. 2814–2822. DOI: 10.1109/CVPRW50498.2020.00338.

[2] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. "VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho. 2022. URL: https://openreview.net/forum?id=AhccnBXSne.

[3] Yuan Gong, Yu-An Chung, and James R. Glass. "AST: Audio Spectrogram Transformer". In: *CoRR* abs/2104.01778 (2021). arXiv: 2104.01778. URL: https://arxiv.org/abs/2104.01778.

[4] Mika Westerlund. "The Emergence of Deepfake Technology: A Review". In: *Technology Innovation Management Review* 9 (Nov. 2019), pp. 39–52. DOI: 10.22215/timreview/1282.

[5] Falko Matern, Christian Riess, and Marc Stamminger. "Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations". In: *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. 2019, pp. 83–92. DOI: 10.1109/WACVW.2019.00020.

[6] Belhassen Bayar and Matthew C. Stamm. "A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer". In: *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*. IH&MMSec '16. Vigo, Galicia, Spain: Association for Computing Machinery, 2016, 5–10. ISBN: 9781450342902. DOI: 10.1145/2909827.2930786. URL: https://doi.org/10.1145/2909827.2930786.

[7] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. "MesoNet: a Compact Facial Video Forgery Detection Network". In: *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. 2018, pp. 1–7. DOI: 10.1109/WIFS.2018.8630761.

[8] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. "Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019, pp. 2307–2311. DOI: 10.1109/ICASSP.2019.8682602.

[9] F. Chollet. "Xception: Deep Learning with Depthwise Separable Convolutions". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, 2017, pp. 1800–1807. DOI: 10.1109/CVPR.2017.195. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.195.

[10] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis. "Two-Stream Neural Networks for Tampered Face Detection". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Los Alamitos, CA, USA: IEEE Computer Society, 2017, pp. 1831–1839. DOI: 10.1109/CVPRW.2017.229. URL: https://doi.ieeecomputersociety.org/10.1109/CVPRW.2017.229.

[11] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. "Not made for each other- Audio-Visual Dissonance-based Deepfake Detection and Localization". In: *Proceedings of the 28th ACM International Conference on Multimedia*. MM '20. Seattle, WA, USA: Association for Computing Machinery, 2020, 439–447. ISBN: 9781450379885. DOI: 10.1145/3394171.3413700. URL: https://doi.org/10.1145/3394171.3413700.

[12] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. "Emotions Don't Lie: An Audio-Visual Deepfake Detection Method using Affective Cues". In: *Proceedings of the 28th ACM International Conference on Multimedia*. MM '20. Seattle, WA, USA: Association for Computing Machinery, 2020, 2823–2832. ISBN: 9781450379885. DOI: 10.1145/3394171.3413570. URL: https://doi.org/10.1145/3394171.3413570.

[13] Yipin Zhou and Ser-Nam Lim. "Joint Audio-Visual Deepfake Detection". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 14780–14789. DOI: 10.1109/ICCV48922.2021.01453.

[14] Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Xiaojun Chang, and Liqiang Nie. "Voice-Face Homogeneity Tells Deepfake". In: *ACM Trans. Multimedia Comput. Commun. Appl.* 20.3 (2023). ISSN: 1551-6857. DOI: 10.1145/3625231. URL: https://doi.org/10.1145/3625231.

[15] Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. "Do You Really Mean That? Content Driven Audio-Visual Deepfake Dataset and Multimodal Method for Temporal Forgery Localization". In: *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*. 2022, pp. 1–10. DOI: 10.1109/DICTA56598.2022.10034605.

[16] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking". In: *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. 2018, pp. 1–7. DOI: 10.1109/WIFS.2018.8630787.

[17] Xin Yang, Yuezun Li, and Siwei Lyu. "Exposing Deep Fakes Using Inconsistent Head Poses". In: *CoRR* abs/1811.00661 (2018). arXiv: 1811.00661. URL: http://arxiv.org/abs/1811.00661.

[18] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, and Lizhuang Ma. "Delving into the Local: Dynamic Inconsistency Learning for DeepFake Video Detection". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36 (June 2022), pp. 744–752. DOI: 10.1609/aaai.v36i1.19955.

[19] Qilin Yin, Wei Lu, Bin Li, and Jiwu Huang. "Dynamic Difference Learning With Spatio–Temporal Correlation for Deepfake Video Detection". In: *IEEE Transactions on Information Forensics and Security* 18 (2023), pp. 4046–4058. DOI: 10.1109/TIFS.2023.3290752.

[20] Ziming Yang, Jian Liang, Yuting Xu, Xiao-Yu Zhang, and Ran He. "Masked Relation Learning for DeepFake Detection". In: *IEEE Transactions on Information Forensics and Security* 18 (2023), pp. 1696–1708. DOI: 10.1109/TIFS.2023.3249566.

[21] Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. *Exploiting Style Latent Flows for Generalizing Deepfake Detection Video Detection*. 2024. arXiv: 2403.06592 [cs.CV].

[22] Davide Cozzolino, Alessandro Pianese, Matthias Nießner, and Luisa Verdoliva. "Audio-Visual Person-of-Interest DeepFake Detection". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2023, pp. 943–952. DOI: 10.1109/CVPRW59228.2023.00101.

[23] Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren. "AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake". In: *IEEE Transactions on Information Forensics and Security* 18 (2023), pp. 2015–2029. DOI: 10.1109/TIFS.2023.3262148.

[24] Shruti Agarwal and Hany Farid. "Detecting Deep-Fake Videos from Aural and Oral Dynamics". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2021, pp. 981–989. DOI: 10.1109/CVPRW53098.2021.00109.

[25] Pavel Korshunov and Sébastien Marcel. *DeepFakes: a New Threat to Face Recognition? Assessment and Detection*. Dec. 2018.

[26] "Google ai blog. contributing data to deepfake detec- tion research. https://ai.googleblog.com/2019 / 09 / contributing - data - to - deepfake - detection.html, 2019." In: (2019).

[27] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. "Celeb-DF: A New Dataset for DeepFake Forensics". In: *CoRR* abs/1909.12962 (2019). arXiv: 1909.12962. URL: http://arxiv.org/abs/1909.12962.

[28] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. "FaceForensics++: Learning to Detect Manipulated Facial Images". In: *CoRR* abs/1901.08971 (2019). arXiv: 1901.08971. URL: http://arxiv.org/abs/1901.08971.

[29] Liming Jiang, Wayne Wu, Ren Li, Chen Qian, and Chen Change Loy. "DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection". In: *CoRR* abs/2001.03024 (2020). arXiv: 2001.03024. URL: http://arxiv.org/abs/2001.03024.

[30] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrer. "The DeepFake Detection Challenge Dataset". In: *CoRR* abs/2006.07397 (2020). arXiv: 2006.07397. URL: https://arxiv.org/abs/2006.07397.

[31] Joel Stehouwer, Hao Dang, Feng Liu, Xiaoming Liu, and Anil K. Jain. "On the Detection of Digital Face Manipulation". In: *CoRR* abs/1910.01717 (2019). arXiv: 1910.01717. URL: http://arxiv.org/abs/1910.01717.

[32] Zhixi Cai, Shreya Ghosh, Aman Pankaj Adatia, Munawar Hayat, Abhinav Dhall, and Kalin Stefanov. *AV-Deepfake1M: A Large-Scale LLM-Driven Audio-Visual Deepfake Dataset*. 2023. arXiv: 2311.15308 [cs.CV].

[33] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. "VoxCeleb2: Deep Speaker Recognition". In: *CoRR* abs/1806.05622 (2018). arXiv: 1806.05622. URL: http://arxiv.org/abs/1806.05622.

[34] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio López-Moreno, and Yonghui Wu. "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis". In: *CoRR* abs/1806.04558 (2018). arXiv: 1806.04558. URL: http://arxiv.org/abs/1806.04558.

[35] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. "A Lip Sync Expert Is All You Need for Speech to Lip Generation In the Wild". In: *Proceedings of the 28th ACM International Conference on Multimedia*. MM '20. Seattle, WA, USA: Association for Computing Machinery, 2020, 484–492. ISBN: 9781450379885. DOI: 10.1145/3394171.3413532. URL: https://doi.org/10.1145/3394171.3413532.

[36] Anurag Bagchi, Jazib Mahmood, Dolton Fernandes, and Ravi Kiran Sarvadevabhatla. "Hear Me Out: Fusional Approaches for Audio Augmented Temporal Action Localiza-

tion". In: *CoRR* abs/2106.14118 (2021). arXiv: 2106.14118. URL: https://arxiv.org/abs/2106.14118.

[37] Chenlin Zhang, Jianxin Wu, and Yin Li. *ActionFormer: Localizing Moments of Actions with Transformers*. 2022. arXiv: 2202.07925 [cs.CV].

[38] Zhixi Cai, Shreya Ghosh, Abhinav Dhall, Tom Gedeon, Kalin Stefanov, and Munawar Hayat. "Glitch in the matrix: A large scale benchmark for content driven audio–visual forgery detection and localization". In: *Computer Vision and Image Understanding* 236 (2023), p. 103818.