TOWARD MULTIMODAL DEEPFAKE DETECTION VIA PHONEME-TEMPORAL AND IDENTITY-DYNAMIC ANALYSIS

by

Tanvi Ranga

January 05, 2026

A dissertation submitted to the
Faculty of the Graduate School of
the University at Buffalo, 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 would like to express my deepest gratitude to my thesis advisor, **Dr. Siwei Lyu**, for giving me the opportunity to work in his research group and for his continuous guidance, encouragement, and support throughout my research. I am also grateful to **Soumyya Kanti Datta** and **Chengzhe Sun**, my co-authors on the publication associated with this thesis, for their valuable collaboration and insights. I sincerely thank **Dr. Nalini Ratha** for serving on my M.S. thesis defense committee and for his thoughtful feedback, which greatly strengthened this work. I would also like to thank my friends for their constant encouragement and unwavering support during this journey. Finally, and most importantly, I extend my heartfelt thanks to my parents and my brother, who have shaped me into the person I am today. I am deeply indebted to them for their love, sacrifices, and unconditional support.

Disclaimer

Parts of this thesis are based on my previously published work at ICCV 2025, Authenticity and Provenance in the age of Generative AI workshop [12], for which I am the first author. This thesis consolidates and extends the contributions from that publication, while providing a unified presentation of the motivation, methodology, and challenges addressed throughout the work. Readers are referred to the above-mentioned paper for additional details. Any overlap between the published work and this thesis reflects my own original contributions as first author.

Table of Contents

List of Tables

List of Figures

Abstract

The rapid advancement of generative media technologies has led to the widespread creation of highly realistic deepfake videos, posing serious risks to identity, trust, and digital security. Existing deepfake detection approaches largely rely on unimodal cues or handcrafted audio-visual alignment rules, which are increasingly ineffective against modern generative models such as Generative Adversarial Networks(GANs) and diffusion models. Although these models generate visually convincing individual frames, they often introduce subtle temporal inconsistencies in speech articulation and facial identity that remain difficult to detect using conventional methods.

To address this challenge, this thesis proposes a novel multimodal audio-visual deepfake detection framework that explicitly models temporal and cross-modal inconsistencies. The proposed model aligns speech and facial motion at the phoneme level and jointly analyzes viseme appearance, lip geometry dynamics, and facial identity embeddings using an attention-based fusion architecture. An auxiliary temporal consistency loss further constrains identity stability across frames, enabling robust detection of face-swap and lip-sync manipulations. Extensive experiments on the FakeAVCeleb and DeepSpeak v2.0 datasets demonstrate that the proposed model achieves state-of-the-art performance and strong generalization, validating the effectiveness of combining phonetic articulation cues with facial motion and identity dynamics for detecting sophisticated deepfakes.

Chapter 1

Introduction

Recent advancements in generative artificial intelligence have facilitated the creation of highly realistic synthetic media. This development has accelerated the creation and spread of deepfake videos that are generated or modified by AI to convincingly modify visual and audio elements. Their growing influence raises serious concerns about authenticity, trust, and digital exploitation.

As generative techniques improve, distinguishing real from manipulated media becomes increasingly challenging for human observers. Visual and audio evidence for authenticity are now easily replicable or disguised. While deepfake technologies have numerous uses in entertainment and media production, their misuse presents considerable threats to human identity, institutional credibility, and societal stability. Occurrences of financial fraud, impersonation, and attacks on social integrity highlight the critical need for effective detection systems.

1.0.1 What Are Deepfakes?

Deepfakes are synthetically manipulated media in which visual or audio components are altered using deep learning techniques to depict events that never occurred. Modern deepfakes are typically produced using deep neural networks, particularly generative models

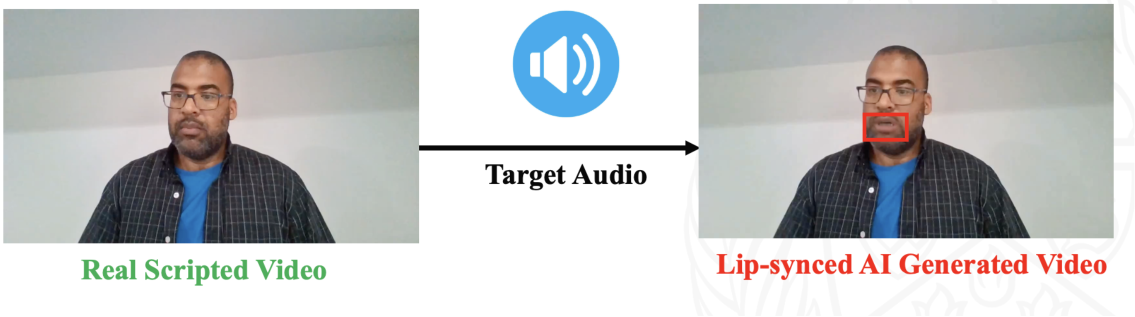

Figure 1.1: Lip-sync deepfake generation with audio-driven mouth synthesis

capable of creating highly realistic facial features, speech, and motion. In audiovisual content, manipulation may affect the visual stream, the audio stream, or both. The two most common forms of human-centered deepfakes are lip-syncing and face-swap.

Lip-Syncing Deepfakes

As illustrated in Figure 1.1, Lip-sync deepfakes modify only the mouth region of a video so that lip movements appear synchronized with a given audio track, while identity, pose, and most facial features remain unchanged. Although modern systems produce smooth and visually convincing results, subtle inconsistencies often persist between phonemes and corresponding articulatory patterns, as well as in the temporal dynamics of mouth motion. These artifacts are difficult for humans to perceive but can be identified through analysis of phoneme articulation, lip geometry, and temporal coherence.

Face-Swap Deepfakes

As illustrated Fig 1.2, in Face-swap deepfakes replace the facial identity of a person in a video with that of another individual while preserving head pose, expressions, and background. Unlike lip-syncing, which affects only the mouth region, face-swapping alters the entire face. While visually realistic, such methods often introduce subtle artifacts related to identity consistency, facial geometry, and temporal stability, including identity drift across frames. These anomalies are more evident when analyzing identity embeddings and tem-

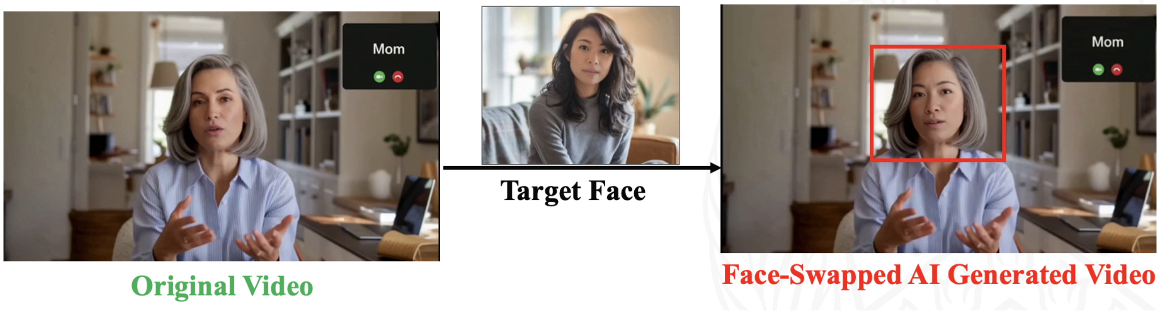

Figure 1.2: Face-swap deepfake generation with identity transfer.

poral coherence rather than individual frames in isolation.

1.0.2 Real-World Implications of Deepfake Misuse

Recent real-world incidents illustrate the significant security, financial, and institutional risks associated with modern deepfake technologies. In February 2024, a multinational organization reportedly suffered a loss of approximately $25 million after an employee was misled by a deepfake impersonation of senior executives and authorized a fraudulent financial transfer. [5] In a separate case, a deepfake impersonator linked to North Korea successfully infiltrated KnowBe4, a cybersecurity firm, by posing as a legitimate employee during the hiring process. [4]

These incidents demonstrate the growing sophistication of audiovisual impersonation techniques and their capacity to exploit organizational trust structures. Beyond financial loss, such attacks undermine institutional credibility, compromise security protocols, and erode public confidence in digital communication. Collectively, these examples highlight the urgent need for reliable and robust deepfake detection systems capable of identifying increasingly subtle and high-fidelity manipulations.

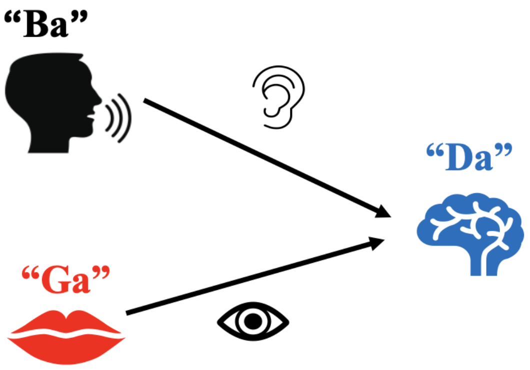

Figure 1.3: Illustration of the McGurk effect

1.1 Foundational Observations

Beyond their societal and economic consequences, the misuse of deepfakes exposes fundamental limitations in current generative models. This thesis proposes a method that builds on empirical evidence that, despite high visual realism, deepfake systems exhibit persistent temporal inconsistencies in articulation and identity preservation. Recognizing these vulnerabilities provides the foundation for the detection framework introduced in this work.

1.1.1 Audiovisual Perception and the McGurk Effect

As shown in Figure 1.3, speech perception is inherently multimodal, relying on the integration of both auditory and visual cues. When the phoneme presented through sound conflicts with the viseme conveyed by lip movements, the brain does not simply favor one modality but instead synthesizes a new percept. This phenomenon demonstrates the strong coupling between phonemes and their corresponding visemic articulations in natural speech. Importantly, it establishes that deviations from this relationship are perceptually meaningful, providing a cognitive basis for detecting inconsistencies between audio and visual speech cues.

1.1.2 Motivation for Temporal Inconsistency Analysis

Motivated by the inherent coupling between auditory and visual speech cues, we hypothesize that deepfake generation models do not consistently reproduce fine-grained visemic articulations, particularly for bilabial and rounded phonemes such as /m/, /b/, /p/, and /o/. In natural speech, these phonemes are associated with distinct and highly repeatable articulatory patterns, including complete lip closure for bilabials and characteristic lip rounding for vowels. Consequently, authentic video exhibits regular, physiologically constrained lip geometry at frames temporally aligned with these sounds.

In contrast, lip-sync and face-manipulation methods often introduce subtle yet systematic deviations, especially during rapid speech, head motion, or complex facial expressions. Although such discrepancies are typically imperceptible at the frame level, they become apparent when examined over time and across modalities. This observation motivates our use of phoneme-conditioned visual analysis and temporal modeling as a principled approach for deepfake detection

Chapter 2

Temporal Articulation and Identity Drift

The proposed solution in this thesis is driven by systematic observations of temporal discrepancies that occur across different types of deepfake videos, such as lip-syncing or face-swap. These anomalies appear as subtle yet measurable discrepancies in audio-visual synchronization and identity preservation over time. These deviations serve as the theoretical foundation for the architectural design decisions and detection methodologies discussed in the next sections of this chapter.

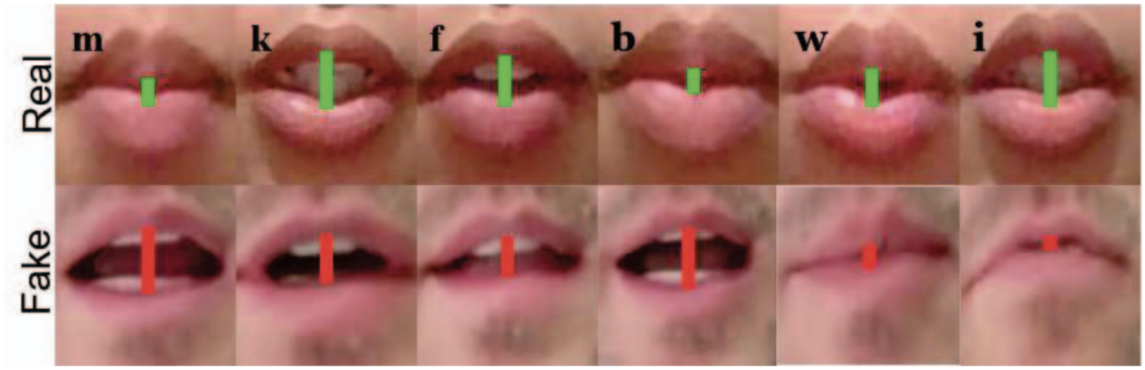

Figure 2.1: Comparison of lip shapes in real and fake video frames corresponding to different phonemes.

Category	**Examples**	**Visual Characteristic**
Bilabials	/p/, /b/, /m/	Complete lip closure, highly visible articulation
Labiodentals	/f/, /v/	Lower lip–upper teeth contact during speech
Alveolars	/t/, /s/	Rapid articulation affecting lip geometry
Velars	/k/	Subtle motion with characteristic mouth shapes
Approximants	/w/, /r/	Lip rounding and forward protrusion
Vowels	/i/, /æ/, /o/	Wide range of mouth openness
Postalveolar	/ʃ/	Distinct lip rounding pattern

Table 2.1: Visually salient phoneme categories used for articulation analysis.

2.1 Phoneme–Viseme Mismatch Patterns

Lip-sync deepfake videos frequently exhibit temporal inconsistencies caused by misalignment between spoken phonemes and the corresponding visual articulations, commonly referred to as visemes. Prior work by Agarwal at el. [1] has shown that accurate phoneme–viseme synchronization is difficult to maintain under generative manipulation, particularly when speech dynamics are complex or rapidly changing. To systematically study this phenomenon, we curate a subset of 14 phonemes selected based on three criteria: the ease of observing lip movements, the diversity of speech gestures, and their significance in identifying deepfakes, as mentioned in Table 2.1.

The selected phonemes span a broad range of phonetic categories, enabling coverage of both high-closure and open-mouth articulations:

This phoneme subset is designed to maximize sensitivity to audiovisual mismatches in manipulated content, where accurate viseme generation is particularly challenging. Phonemes with limited visual discriminability, such as glottal sounds or unstressed vowels, as well as silence segments, are excluded to reduce noise and avoid uninformative supervision sig-

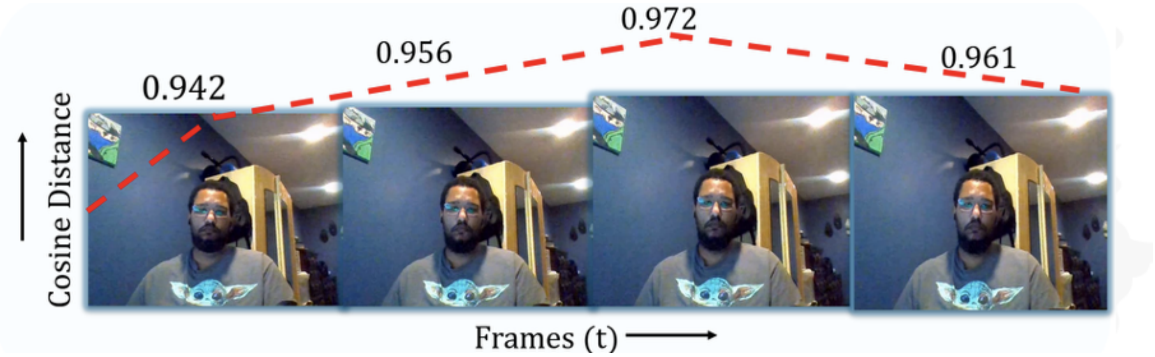

Figure 2.2: Cosine distance of ArcFace identity embeddings in face-swap videos

nals.

Through empirical analysis, we find that lip-sync manipulation methods often fail to maintain consistent alignment between phoneme articulation and lip motion, particularly for phonemes requiring precise closure or controlled openness, such as /p/, /b/, /m/, /f/, /v/, and /o/. In many manipulated sequences, the observed lip geometry deviates from the articulatory configuration implied by the audio. As shown in Figure 2.1, lip-sync videos exhibit clear mismatches between expected and observed mouth shapes, especially during high-closure phoneme events. These discrepancies are reliably captured using MediaPipe-based lip landmark analysis, which measures lip closure, aspect ratio, and mouth openness at the frame level, and their persistence over time makes them a robust detection cue.

2.2 Temporal Drift in Identity Embeddings

Beyond phoneme–viseme misalignment, we observe temporal inconsistencies in facial identity representations extracted from manipulated videos. Identity embeddings are computed using an ArcFace-based face recognition model and analyzed across consecutive frames to evaluate identity preservation over time. For AI-generated face-swap videos, the cosine distance in Arcface embeddings is inconsistent and shows high spikes as shown in Fig 2.2. This phenomenon is further quantified in Fig 2.3, which plots the *L2* distance between ArcFace embeddings of consecutive frames for both real and manipulated videos.

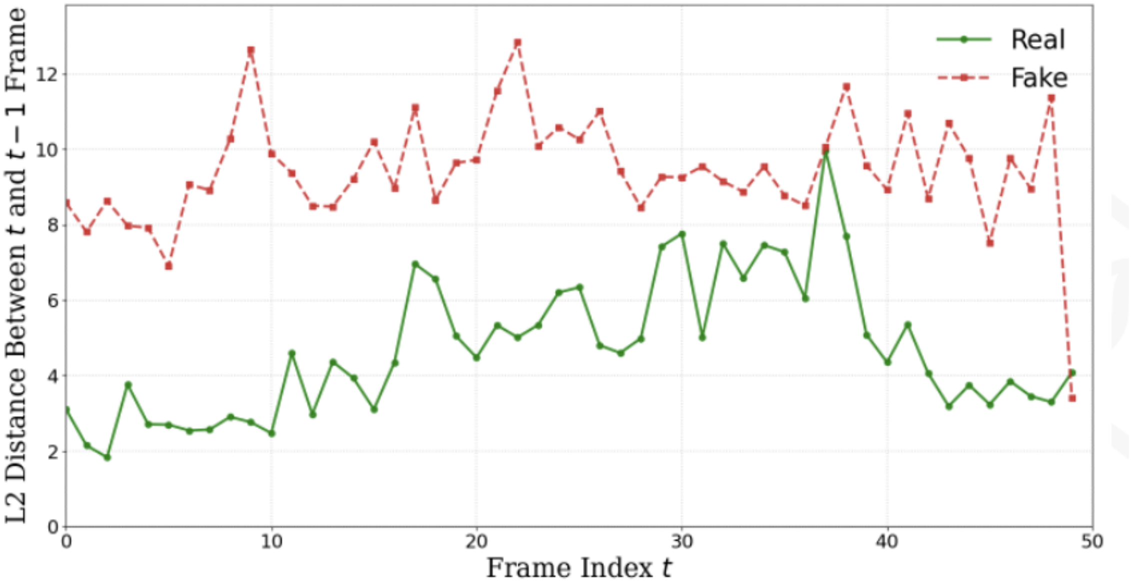

Figure 2.3: ArcFace L2 drift for real and fake video frames.

Real videos exhibit consistently low embedding distances, reflecting gradual and continuous identity transitions. In contrast, face-swap videos demonstrate persistently higher and more volatile embedding distances, indicating temporal identity drift. Such behavior arises from imperfect face blending, pose-dependent synthesis artifacts, and inconsistent feature reconstruction across frames.

These findings show that temporal consistency in identity embeddings provides a strong cue for face-swap detection, enabling reliable separation of real and manipulated videos.

Chapter 3

Related Work

3.1 Deepfake Generation

Advances in artificial intelligence have enabled increasingly realistic and accessible deepfake generation, broadly categorized into full-face synthesis (e.g., face-swaps and talking-head generation) and partial manipulation, most commonly lip-syncing. Early work focused on face replacement [22, 28, 35, 36, 6], while lip-syncing methods such as Wav2Lip [32] and VideoReTalking [9] modify only the mouth region to match speech. More recent diffusion-based approaches [27, 23] improve visual fidelity and temporal coherence, and avatar-based systems [16, 40, 41] animate a single image into a talking head using identity, motion, and speech embeddings, further increasing realism and making detection more challenging.

3.2 Deepfake Detection

As generation quality improves, detection methods have evolved accordingly. Visual-only approaches analyze spatial artifacts and temporal inconsistencies [34, 24, 29, 19, 11, 18, 42], while audio–visual methods exploit mismatches between speech and lip motion [14, 39, 38]. Representative works include phoneme–viseme mismatch detection [1], multimodal fusion of audio, visual, and physiological cues [7], self-supervised learning of audio–visual correspondences [30], and recent transformer-based models that capture fine-grained temporal inconsistencies using ViTs [13] and CLIP features [33] [25].

3.3 Unaddressed Challenges in Deepfake Detection

Despite significant progress, many detectors rely on limited cues such as pixel artifacts, coarse audio–visual synchronization, or unimodal temporal patterns. As modern generative models increasingly produce visually coherent frames with improved lip synchronization and identity consistency, these cues become less reliable, particularly under cross-manipulation and cross-dataset settings. This motivates the need for more comprehensive detection frameworks that explicitly integrate multimodal information and model fine-grained temporal dynamics.

Chapter 4

Proposed Method

As discussed in previous chapters, existing deepfake detectors are largely limited by rule-based heuristics or unimodal designs, which are unable to capture the subtle cross-modal and temporal inconsistencies introduced by modern face-swap and lip-sync techniques. To overcome these limitations, this thesis proposes *PIA*, a two-stage multimodal framework that jointly models phoneme articulation, lip geometry, viseme appearance, and facial identity consistency over time. [12]

Figure 4.1 illustrates the end-to-end architecture of our proposed solution. The first stage performs multimodal feature extraction and temporal alignment across audio, visual, geometric, and identity embeddings. The second stage integrates these representations using multi-head attention-based fusion and temporal pooling to identify subtle anomalies by correlating phoneme-level audio with visual mouth motion. Visual features are extracted

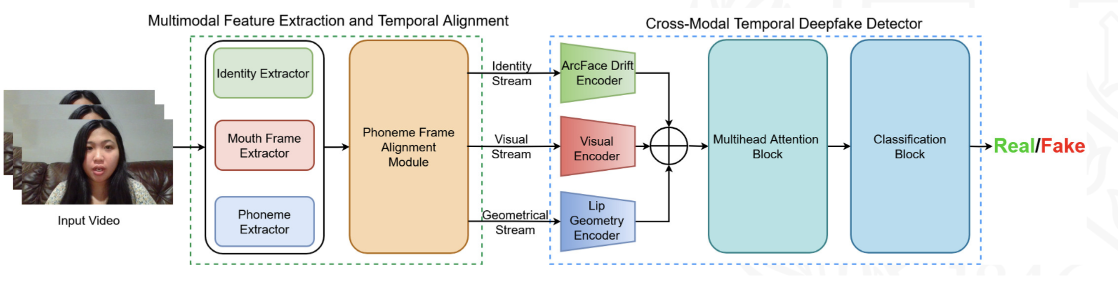

Figure 4.1: Proposed PIA model.

using a 3D convolutional neural network(CNN) followed by a pretrained EfficientNet-B0, while temporal dependencies and modality interactions are modeled using multi-head attention. Phonemes extracted via WhisperX [2] and aligned with wav2vec2 are used as an active filter to select frames with visually meaningful articulation.

4.1 Feature Extraction

Each video is processed through a structured multi-stage preprocessing pipeline designed to extract synchronized multimodal information. This pipeline includes audio extraction and phoneme alignment to obtain precise speech–time correspondence, viseme feature extraction to capture mouth-region appearance and articulatory motion, facial identity embedding to model identity consistency across frames, and multimodal frame-level alignment to ensure temporal correspondence between all extracted audio, visual, geometric, and identity features.

4.1.1 Audio Extraction and Phoneme Alignment

Audio is extracted using FFmpeg[15], resampled to 16 kHz mono, and transcribed using WhisperX[2]. Word-level segments are converted to IPA phonemes using the phonemizer toolkit. Later, a wav2vec2 aligner is employed for precise frame-level phoneme labeling. Each frame is assigned a phoneme label by timestamp interpolation, ensuring precise audio–visual synchronization.

4.1.2 Viseme Feature Extraction

MediaPipe FaceMesh detects 468 landmarks per frame. From the detected facial landmarks, 27 points corresponding to the mouth region are selected, and geometric descriptors are derived. Lip height and lip width are subsequently used to compute the mouth aspect ratio. Lip height and width are used to compute the mouth aspect ratio (MAR):

$$\mathrm{MAR} = \frac{\text{lip\_height}}{\text{lip\_width} + \varepsilon}, \tag{4.1}$$

where ε ensures numerical stability. MAR encodes mouth openness, with low values for bilabials (/m/, /b/, /p/) and higher values for open vowels, capturing phoneme-specific articulation and revealing manipulation inconsistencies.

4.1.3 Facial Identity Embedding

Frame-level identity embeddings are extracted using ArcFace, which maps faces into a 512-dimensional hyperspherical space with strong inter-class separability. These embeddings capture expression-invariant facial structure and enable detection of identity drift across frames. Identity features are used both as model inputs and for auxiliary temporal consistency regularization.

4.1.4 Multimodal Representation Construction

A phoneme-aligned dataset is built using 14 visually distinct phonemes. For each phoneme instance, five frames are sampled. Each frame yields three modalities: mouth-region viseme crops, ArcFace embeddings, and geometric lip descriptors. Non-linguistic segments are excluded, focusing training on visually informative articulation.

4.2 Model Architecture

The proposed framework adopts a multistream architecture that jointly models three complementary modalities: lip geometry, viseme appearance, and facial identity. Each modality captures a distinct aspect of audiovisual consistency and is processed by a dedicated encoder prior to fusion. For each phoneme instance, the first 5 consecutive frames are sampled. At each time step, the input consists of a mouth-region crop representing visual appearance, a scalar lip geometry descriptor (MAR), and a 512-dimensional ArcFace embedding encoding facial identity. The objective of the architecture is to determine whether these modality-specific signals remain mutually consistent over time, as expected in authentic videos, or whether they exhibit subtle mismatches characteristic of lip-sync and face-swap manipulations.

4.2.1 Visual Encoder

The visual stream is designed to model short-term mouth dynamics and local appearance artifacts within the lip region. The T mouth-region crops are stacked to form a spatiotemporal clip and processed by a three-dimensional convolutional neural network (3D CNN), which jointly learns spatial and temporal features. Unlike frame-wise 2D CNNs, the 3D CNN explicitly captures motion-sensitive patterns such as lip opening and closure trajectories, temporal smoothness, and frame-to-frame texture consistency, all of which are often disrupted in manipulated content.

The resulting feature maps are aggregated across time using temporal averaging to obtain a compact and noise-robust representation of phoneme-level mouth motion. This aggregated feature is subsequently passed through a pretrained EfficientNet-B0 backbone to extract higher-level visual representations. EfficientNet-B0 provides strong generalization and enables the encoder to capture mid-level and semantic features related to edge structure, texture regularity, and regional coherence. The output of this stream is a d-dimensional visual embedding $\mathbf{v}_t \in \mathbb{R}^d$ that summarizes both spatial appearance and short-term temporal behavior of the mouth region.

4.2.2 Lip Geometry Encoder

The lip geometry stream models articulatory constraints that are difficult for generative models to reproduce consistently over time. For each frame, the mouth aspect ratio (MAR) is computed as a scalar measure of mouth openness. Over a phoneme window, the resulting MAR sequence forms a low-dimensional temporal signal describing how the mouth opens and closes during speech.

This T-dimensional MAR sequence is passed through a lightweight multilayer perceptron (MLP) to produce a geometry embedding $\mathbf{g}_t \in \mathbb{R}^d$.

4.2.3 Identity Encoder

The identity stream models the temporal consistency of facial identity. For each frame, a 512-dimensional ArcFace embedding is extracted, encoding expression-invariant facial attributes. To reduce noise from pose, blur, and occlusion, embeddings are aggregated across the phoneme window and passed through an MLP to produce the identity representation $\mathbf{a}_t \in \mathbb{R}^d$. This branch is particularly effective for face-swap detection, as manipulated videos often exhibit subtle identity drift over time despite appearing visually plausible frame by frame.

4.2.4 Multimodal Fusion and Temporal Attention

At each time step t, the modality-specific embeddings from the geometry, visual, and identity streams are fused by concatenation:

$$\mathbf{f}_t = \mathbf{g}_t \oplus \mathbf{v}_t \oplus \mathbf{a}_t \in \mathbb{R}^{3d}, \tag{4.2}$$

where $\mathbf{g}_t$, $\mathbf{v}_t$, and $\mathbf{a}_t$ denote the geometry, visual, and identity embeddings, respectively. The fused vector $\mathbf{f}_t$ jointly encodes articulatory motion, visual appearance, and identity consistency for the corresponding phoneme instance. The sequence $\{\mathbf{f}_t\}_{t=1}^{T}$ therefore represents the temporal evolution of these multimodal cues.

To aggregate this sequence into a single video-level representation, multi-head attention pooling is employed. Unlike uniform averaging, attention pooling allows the model to assign higher weights to temporally informative frames or phoneme instances, reflecting the fact that manipulation artifacts are often localized in time and that certain phonemes are more visually diagnostic. For each attention head h, a set of normalized attention weights $\{\alpha_{h,t}\}_{t=1}^{T}$ is learned such that $\sum_{t=1}^{T} \alpha_{h,t} = 1$. The pooled representation is computed as:

$$\mathbf{z} = \frac{1}{H} \sum_{h=1}^{H} \sum_{t=1}^{T} \alpha_{h,t} \mathbf{f}_t, \tag{4.3}$$

where H denotes the number of attention heads. The resulting vector $\mathbf{z}$ summarizes the most informative multimodal evidence across time and is passed to a fully connected classifier to predict whether the input video is real or manipulated.

4.3 ArcFace Temporal Consistency Loss

While the classification objective encourages separation between real and fake samples, it does not explicitly constrain identity features to evolve smoothly over time. To regularize identity dynamics, a temporal consistency loss is introduced on the ArcFace embeddings. For each pair of consecutive frames, cosine similarity is computed as:

$$s_t = \cos(\mathbf{a}_t, \mathbf{a}_{t+1}) = \frac{\mathbf{a}_t^{\top} \mathbf{a}_{t+1}}{\|\mathbf{a}_t\|_2 \, \|\mathbf{a}_{t+1}\|_2}. \tag{4.4}$$

where $\mathbf{a}_t$ and $\mathbf{a}_{t+1}$ denote identity embeddings of adjacent frames. High values of s_t indicate stable identity across frames, while lower values indicate identity drift.

To avoid penalizing unreliable segments such as silence, occlusion, or poor face detection, a binary mask $m_t \in \{0, 1\}$ is applied. The temporal consistency loss is defined as:

$$\mathcal{L}_{\text{arcface}} = \frac{\sum_{t=1}^{T-1} (1 - s_t) \, m_t \, m_{t+1}}{\sum_{t=1}^{T-1} m_t \, m_{t+1} + \epsilon}, \tag{4.5}$$

where ϵ is a small constant for numerical stability. This loss penalizes large identity changes between consecutive frames in valid regions, encouraging smooth identity evolution in genuine videos while amplifying the effect of identity drift commonly observed in face-swap deepfakes.

4.3.1 Overall Training Objective

The final training objective combines the standard cross-entropy classification loss with the identity regularization term:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{arcface}}, \tag{4.6}$$

where λ controls the contribution of the temporal consistency constraint. This joint objective enables robust multimodal learning while explicitly modeling identity stability, thereby improving discrimination between authentic and manipulated videos.

Chapter 5

Experiments

This chapter presents the experimental setup used to evaluate the proposed framework, including the datasets, data splits, implementation details, and evaluation metrics. All experiments are designed to assess performance across diverse manipulation types.

5.1 Experimental Settings

5.1.1 Datasets

Experiments are conducted on two public benchmarks: FakeAVCeleb [21] and DeepSpeak v2.0 [3]. Both datasets contain face-swap, lip-sync, and avatar-based forgeries, making them suitable for evaluating multimodal deepfake detection. Sample frames from both datasets are shown in Fig. 5.1.

FakeAVCeleb

The FakeAVCeleb dataset [21] consists of 20,000 videos (19,500 fake and 500 real) at a resolution of 224×224. Following the protocol in [31], the data are grouped into five categories: FVRA-WL (Wav2Lip), FVFA-FS (FaceSwap), FVFA-GAN (FaceSwapGAN), FVFA-WL (Wav2Lip), and RVFA (Real Video, Fake Audio). A 70:30 train–test split is

FakeAVCeleb [*H Khalid et al.*, 2021]

Deepspeak v2 [*S Barrington, et al.*, 2024]

Figure 5.1: Samples from the evaluated datasets.

adopted as in [31]. The RVFA category is excluded from both training and testing in order to focus specifically on video-level manipulation.

DeepSpeak v2.0

The DeepSpeak v2.0 dataset [3] contains 9,376 real and 7,209 fake videos covering face-swap, lip-sync, and avatar-based forgeries, with resolutions ranging from 640×480 to 1280×720. The official 80:20 train–test split is followed. The test set is further partitioned by manipulation type to evaluate robustness across different forgery methods.

5.1.2 Implementation Details

The model is implemented in PyTorch 2.6.0 with CUDA 12.4. Fourteen visually distinct phonemes are extracted using WhisperX [2] and temporally aligned to video frames. Mouth-region crops are resized to 112×112 and normalized prior to input. Training is performed using cross-entropy loss with label smoothing, along with an auxiliary ArcFace temporal consistency loss. Optimization is carried out using Adam with a learning rate of 3×10^{-4}, weight decay of 1×10^{-5}, $\lambda = 0.1$ for the auxiliary loss, 4 attention heads, 25 training epochs, and a batch size of 16.

5.1.3 Evaluation Metrics

Performance is evaluated using Accuracy (ACC), Area Under the ROC Curve (AUC), and Average Precision (AP). All results are reported in percentage points (%-pts) to ensure consistent comparison across datasets and manipulation types.

Chapter 6

Results and Discussion

This chapter presents the experimental results of the proposed PIA framework and analyzes its behavior across datasets and manipulation types. Quantitative results on FakeAVCeleb are first reported, including a cross-manipulation evaluation. Results on DeepSpeak v2.0 are then presented, followed by a summary of findings from an ablation study.

6.1 Results on FakeAVCeleb

6.1.1 Overall Performance

Following the evaluation protocol in [31], training is performed on the FakeAVCeleb [21] training split, and evaluation is conducted on the official test split. Table 6.1 compares PIA with visual-only and audiovisual baselines using ACC and AUC.

The proposed framework achieves the highest overall performance, reaching 98.7% ACC and 99.8% AUC, outperforming all baselines, including AVFF [31]. Results are also reported for PIA_RVFA, where the RVFA category is introduced only during testing, while remaining excluded from training. Under this setting, performance of 98.0% ACC and 98.2% AUC is obtained, indicating that the learned representations remain effective even when audio manipulation is present at test time.

Method	Modality	ACC(%)	AUC(%)
Xception [34]	V	67.9	70.5
LipForensics [18]	V	80.1	82.4
FTCN [42]	V	64.9	84.0
CViT [37]	V	69.7	71.8
RealForensics [17]	V	89.9	94.6
Emotions Don't Lie [26]	AV	78.1	79.8
MDS [10]	AV	82.8	86.5
AVFakeNet [20]	AV	78.4	83.4
VFD [8]	AV	81.5	86.1
AVoID-DF [38]	AV	83.7	89.2
AVFF [31]	AV	98.6	99.1
PIA_RVFA (Ours)	AV	98.0	98.2
PIA (Ours)	AV	**98.7**	**99.8**

Table 6.1: Performance on FakeAVCeleb.

6.1.2 Cross Manipulation Generalization

Generalization to unseen manipulation methods is evaluated following the cross-manipulation protocol in [31]. Using the four categories (FVRA-WL, FVFA-FS, FVFA-GAN, FVFA-WL), training is conducted on three categories, and testing is performed on the held-out category, with this procedure repeated for all categories. Table 6.2 reports AP and AUC results.

Across all held-out settings, the proposed model, PIA, achieves the strongest performance. On FVRA-WL, AP is improved over LipForensics [18] by 2.1 percentage points, and AUC is improved over AVFF [31] by 0.9 percentage points. Averaged across all held-out conditions (AVG-FV), performance exceeds AVFF by 1.4 points in AP and 0.3 points in AUC. These results indicate that transferable manipulation cues are learned rather than method-specific artifacts.

Method	Modality	FVRA–WL		FVFA–FS		FVFA–GAN		FVFA–WL		AVG–FV	
		AP (%)	AUC (%)	AP (%)	AUC (%)	AP (%)	AUC (%)	AP (%)	AUC (%)	AP (%)	AUC (%)
Xception [34]	V	88.2	88.3	92.3	93.5	67.6	68.5	91.0	91.0	84.8	85.3
LipForensics [18]	V	97.8	97.7	99.9	99.9	61.5	68.1	98.6	98.7	89.4	91.1
FTCN [42]	V	96.2	97.4	**100.0**	**100.0**	77.4	78.3	95.6	96.5	92.3	93.1
RealForensics [17]	V	88.8	93.0	99.3	99.1	99.8	99.8	93.4	96.7	95.3	97.1
AV-DFD [43]	AV	97.0	97.4	99.6	99.7	58.4	55.4	**100.0**	**100.0**	88.8	88.1
AVAD (LRS2) [14]	AV	93.6	93.7	95.3	95.8	94.1	94.3	93.8	94.1	94.2	94.5
AVAD (LRS3) [14]	AV	91.1	93.0	91.0	92.3	91.6	92.7	91.4	93.1	91.3	92.8
AVFF [31]	AV	94.8	98.2	**100.0**	**100.0**	99.9	**100.0**	99.4	99.8	98.5	99.5
PIA (Ours)	AV	**99.9**	**99.1**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**99.9**	**99.8**

Table 6.2: Cross manipulation evaluation on FakeAVCeleb.

6.2 Results on DeepSpeak v2.0

The DeepSpeak v2.0 dataset [3] includes higher-quality videos and avatar-based deepfakes. Training and evaluation are performed using the official train–test split. Table 6.3 reports AUC performance on the lip-sync, face-swap, avatar, and combined global test sets.

Strong performance is achieved across all manipulation types, with a global AUC of 98.06%. These results demonstrate that the proposed framework remains effective for high-fidelity manipulations and avatar-based generation, which exhibit different visual characteristics from conventional face-swap and lip-sync deepfakes.

Dataset	Lip-sync	Face-swap	Avatar	Global
PIA_w_ph_w/o_vi	68.44	62.12	64.73	65.57
PIA_w_ph_w/o_geom	98.31	91.56	96.77	96.49
PIA_w_ph_w/o_arc	98.64	95.99	96.68	97.02
PIA_w_ph	98.95	92.54	96.43	96.66
PIA_w/o_EB0	94.81	81.70	86.54	88.68
Vgg16_w/o_PIA	91.51	78.36	85.49	86.62
PIA	**99.24**	**96.47**	**97.76**	**98.06**

Table 6.3: Ablation analysis on Deepspeak v2.0 test set

6.3 Ablation Summary

An ablation study is reported in Table 6.3 to evaluate the contribution of each component of the proposed PIA framework on the DeepSpeak v2.0 dataset.

- **Excluding Visemes (PIA w ph w/o vi):** Removing viseme image embeddings results in the largest degradation, with AUC drops of 30.8, 34.35, 33.03, and 32.49 percentage points on the Lip-sync, Face-swap, Avatar, and Global subsets, respectively, confirming mouth appearance as the most discriminative cue.

- **Excluding Lip Geometry (PIA w ph w/o geom):** Eliminating the geometry stream leads to smaller but consistent reductions in AUC of 0.93, 4.91, 0.99, and 1.57 points across the Lip-sync, Face-swap, Avatar, and Global subsets, indicating that articulatory geometry provides complementary information.

- **Excluding ArcFace Embeddings (PIA w ph w/o arc):** Removing identity features causes modest yet systematic drops in AUC of 0.60, 0.48, 1.08, and 1.04 points, highlighting the importance of temporal identity stability, particularly for face-swap detection.

- **Including One-Hot Phonemes (PIA w ph):** Direct fusion of one-hot phoneme features slightly degrades performance, with AUC decreases of 0.29, 3.93, 1.33, and 1.40 points, suggesting that phonemes are more effective for alignment and frame selection than as explicit fusion inputs.

- **Excluding EfficientNet-B0 (PIA w/o EB0):** Replacing the EfficientNet-B0 backbone with frozen ResNet-18 features results in substantial AUC drops of 4.43, 14.77, 11.22, and 9.38 points, demonstrating the necessity of a strong, trainable visual backbone.

- **Using VGG16 (Vgg16 w/o PIA):** Substituting the PIA architecture with a simpler VGG16 model causes significant losses of 7.73, 18.11, 12.27, and 11.44 AUC points across subsets, underscoring the value of jointly modeling viseme appearance, lip geometry, and identity dynamics.

Overall, the results confirm that integrating visual, geometric, and identity cues within the full PIA architecture yields the most robust and generalizable deepfake detection performance.

6.4 Summary of Findings

Across both FakeAVCeleb and DeepSpeak v2.0, strong performance and robust generalization to unseen manipulation methods are observed. The results highlight the importance of combining viseme appearance with temporal modeling, while lip geometry and identity stability provide complementary gains, particularly for challenging manipulation types.

Chapter 7

Conclusion

This thesis introduced PIA (Phoneme-Temporal and Identity-Dynamic Analysis), a multimodal framework for audiovisual deepfake detection that targets the subtle temporal and cross-modal inconsistencies and address the current gap in unimodal or rule based methods. PIA jointly models phoneme-aligned articulation, mouth appearance, lip geometry, and identity dynamics through an end-to-end alignment pipeline, a multistream architecture with attention-based fusion, and an ArcFace temporal consistency loss to capture identity drift in face-swap manipulations. Experiments demonstrate strong performance and generalization, achieving state-of-the-art results on FakeAVCeleb and consistently high accuracy on DeepSpeak v2.0, including avatar-based deepfakes. Remaining limitations include reduced robustness to unseen resolutions, reliance on English-based WhisperX and wav2vec2 alignment, and the lack of explicit handling for Real Video Fake Audio (RVFA). Future work will address these through improved resolution robustness, RVFA modeling, and multilingual, language-agnostic speech representations, strengthening PIA for real-world deployment.

Biblography

[1] Shruti Agarwal et al. “Detecting deep-fake videos from phoneme-viseme mismatches”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020, pp. 660–661.

[2] Max Bain et al. “WhisperX: Time-Accurate Speech Transcription of Long-Form Audio”. In: *INTERSPEECH 2023* (2023).

[3] Sarah Barrington, Matyas Bohacek, and Hany Farid. “DeepSpeak Dataset v1. 0”. In: *arXiv preprint arXiv:2408.05366* (2024).

[4] Matt Bracken. “Cyber firm KnowBe4 hired a fake IT worker from North Korea”. In: *CyberScoop* (July 2024). URL: `https://cyberscoop.com/cyber-firm-knowbe4-hired-a-fake-it-worker-from-north-korea/`.

[5] Heather Chen and Kathleen Magramo. *Finance worker pays out $25 million after video call with deepfake ‘chief financial officer’*. https://www.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html. Published by CNN; accessed 30-January-2025.

[6] Renwang Chen et al. “Simswap: An efficient framework for high fidelity face swapping”. In: *Proceedings of the 28th ACM international conference on multimedia*. 2020, pp. 2003–2011.

[7] Yu Chen et al. “Npvforensics: Jointing non-critical phonemes and visemes for deepfake detection”. In: *arXiv preprint arXiv:2306.06885* (2023).

[8] Harry Cheng et al. “Voice-face homogeneity tells deepfake”. In: *ACM Transactions on Multimedia Computing, Communications and Applications* 20.3 (2023), pp. 1–22.

[9] Kun Cheng et al. “Videoretalking: Audio-based lip synchronization for talking head video editing in the wild”. In: *SIGGRAPH Asia 2022 Conference Papers*. 2022, pp. 1–9.

[10] Komal Chugh et al. “Not made for each other-audio-visual dissonance-based deepfake detection and localization”. In: *Proceedings of the 28th ACM international conference on multimedia*. 2020, pp. 439–447.

[11] Soumyya Kanti Datta, Shan Jia, and Siwei Lyu. “Exposing lip-syncing deepfakes from mouth inconsistencies”. In: *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2024, pp. 1–6.

[12] Soumyya Kanti Datta et al. “PIA: Deepfake Detection Using Phoneme-Temporal and Identity-Dynamic Analysis”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*. Oct. 2025, pp. 1596–1606.

[13] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[14] Chao Feng, Ziyang Chen, and Andrew Owens. “Self-supervised video forensics by audio-visual anomaly detection”. In: *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 10491–10503.

[15] FFmpeg Developers. *FFmpeg*. https://ffmpeg.org/. Version 4.4.2. (Visited on 08/09/2025).

[16] Jianzhu Guo et al. “Liveportrait: Efficient portrait animation with stitching and retargeting control”. In: *arXiv preprint arXiv:2407.03168* (2024).

[17] Alexandros Haliassos et al. “Leveraging real talking faces via self-supervision for robust forgery detection”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 14950–14962.

[18] Alexandros Haliassos et al. “Lips don’t lie: A generalisable and robust approach to face forgery detection”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 5039–5049.

[19] Baojin Huang et al. “Implicit identity driven deepfake face swapping detection”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 4490–4499.

[20] Hafsa Ilyas, Ali Javed, and Khalid Mahmood Malik. “AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio-visual deepfakes detection”. In: *Applied Soft Computing* 136 (2023), p. 110124.

[21] Hasam Khalid et al. “FakeAVCeleb: A novel audio-video multimodal deepfake dataset”. In: *arXiv preprint arXiv:2108.05080* (2021).

[22] Iryna Korshunova et al. “Fast face-swap using convolutional neural networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3677–3685.

[23] Chunyu Li et al. “LatentSync: Audio Conditioned Latent Diffusion Models for Lip Sync”. In: *arXiv preprint arXiv:2412.09262* (2024).

[24] Yuezun Li and Siwei Lyu. “Exposing deepfake videos by detecting face warping artifacts”. In: *arXiv preprint arXiv:1811.00656* (2018).

[25] Weifeng Liu et al. “Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 91131–91155.

[26] Trisha Mittal et al. “Emotions don’t lie: An audio-visual deepfake detection method using affective cues”. In: *Proceedings of the 28th ACM international conference on multimedia*. 2020, pp. 2823–2832.

[27] Soumik Mukhopadhyay et al. “Diff2lip: Audio conditioned diffusion models for lip-synchronization”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 5292–5302.

[28] Yuval Nirkin, Yosi Keller, and Tal Hassner. “Fsgan: Subject agnostic face swapping and reenactment”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 7184–7193.

[29] Yuval Nirkin et al. “Deepfake detection based on discrepancies between faces and their context”. In: *IEEE transactions on pattern analysis and machine intelligence* 44.10 (2021), pp. 6111–6121.

[30] Trevine Oorloff et al. “Avff: Audio-visual feature fusion for video deepfake detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 27102–27112.

[31] Trevine Oorloff et al. “Avff: Audio-visual feature fusion for video deepfake detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 27102–27112.

[32] KR Prajwal et al. “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*. 2020, pp. 484–492.

[33] Alec Radford et al. “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PmLR. 2021, pp. 8748–8763.

[34] Andreas Rossler et al. "Faceforensics++: Learning to detect manipulated facial images". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 1–11.

[35] Henry Ruhs. *Facefusion*. 2024. URL: `https://github.com/facefusion/facefusion`.

[36] Haofan Wang. *INSwapper: Face swapping model based on insightface*. 2023. URL: `https://github.com/haofanwang/inswapper`.

[37] Deressa Wodajo and Solomon Atnafu. "Deepfake video detection using convolutional vision transformer". In: *arXiv preprint arXiv:2102.11126* (2021).

[38] Wenyuan Yang et al. "Avoid-df: Audio-visual joint learning for detecting deepfake". In: *IEEE Transactions on Information Forensics and Security* 18 (2023), pp. 2015–2029.

[39] Cai Yu et al. "Explicit correlation learning for generalizable cross-modal deepfake detection". In: *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2024, pp. 1–6.

[40] Shengkai Zhang et al. "HelloMeme: Integrating Spatial Knitting Attentions to Embed High-Level and Fidelity-Rich Conditions in Diffusion Models". In: *arXiv preprint arXiv:2410.22901* (2024).

[41] Longtao Zheng et al. "MEMO: Memory-Guided Diffusion for Expressive Talking Video Generation". In: *arXiv preprint arXiv:2412.04448* (2024).

[42] Yinglin Zheng et al. "Exploring temporal coherence for more general video face forgery detection". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 15044–15054.

[43] Yipin Zhou and Ser-Nam Lim. "Joint audio-visual deepfake detection". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 14800–14809.