Robust Cross-Modal Deepfake Detection via Facial UV Maps and Momentum Contrastive Learning

Yuesen Tang[1], Yuanyang Zhang[1], Wangxiao Mao[1], Li Yao[1†]
{22046380, 22046331, zhangyuanyang, yao.li}@seu.edu.cn
[1]School of Computer Science and Engineering, Southeast University, Nanjing, China
[†]Corresponding Author

***Abstract*—With the rapid advancement of deepfake technology, its application in multimodal content such as video and audio has posed severe threats to digital information security and privacy protection. Although existing deepfake detection methods have made significant progress, three critical challenges remain: efficient fusion of cross-modal information, precise extraction of fine-grained visual features, and effective discrimination between the distributions of authentic and forged samples. To address these challenges, this paper innovatively proposes FUME, a cross-modal deepfake detection framework integrating facial UV maps with momentum contrastive learning. Specifically, the method introduces a Texture-Aware Video Transformer (TAViT) to achieve deep fusion of facial UV maps and spatiotemporal video features, while employing an Audio Spectrogram Transformer (AST) for multi-scale feature modeling of speech signals. Additionally, we design a momentum contrastive learning based cross-modal alignment mechanism, which achieves semantic-level matching of audiovisual representations through dynamic construction of positive and negative sample pairs, while incorporating a One-Class softmax loss function to enhance generalization capability against unseen deepfake generation techniques. Extensive experiments on datasets including DeepfakeTIMIT, DFDC, and KoDF validate our method's superiority. Notably, on the KoDF dataset, our approach achieves 97.59% accuracy and 98.23% AUC, surpassing the current audio-visual state-of-the-art by 1.91% and 2.99%.**

***Index Terms*—Cross-modal learning, deepfake detection, contrastive learning, facial UV maps.**

I. Introduction

In recent years, the rapid development and proliferation of deepfake technology have garnered substantial attention across a multitude of fields, primarily due to its remarkable capacity to generate highly realistic synthetic media content. Leveraging the sophisticated capabilities of artificial intelligence and deep learning algorithms, deepfake methodologies are now able to fabricate images and videos with such a high degree of photorealism that they are often virtually indistinguishable from genuine, unaltered material [1]. This technological advancement has enabled a broad spectrum of applications, ranging from film and digital entertainment to political communication, interpersonal social interactions, and even security operations. Despite its innovative potential and utility in creative and commercial contexts, deepfake technology has simultaneously provoked serious ethical and societal concerns, particularly with regard to individual privacy infringement and the widespread dissemination of false or misleading information [2]. What initially emerged as a novel tool within the realm of the entertainment industry has, through rapid technological progression, transformed into a formidable instrument with the power to influence public perception, manipulate discourse, and pose significant threats to societal trust and stability [3], [4].

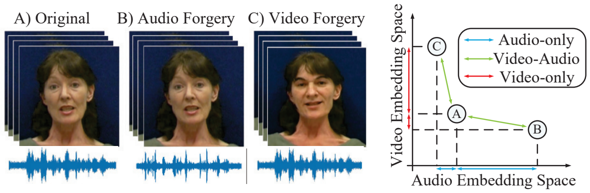

Fig. 1. Multimodal Embedding Space for Differentiating Audio and Video Forgeries

Deepfake detection methods can be broadly classified into two main categories: unimodal and multimodal approaches. Unimodal detection techniques concentrate on analyzing a single modality—either visual or auditory—to identify anomalies that may indicate manipulated content. In the visual domain, these methods typically detect inconsistencies in facial expressions, unnatural eye blinking, or irregular head movements within video frames, while in the auditory domain, they often examine disruptions in pitch, timbre, or background noise to uncover traces of audio tampering [5], [6]. Although unimodal methods have demonstrated effectiveness in controlled scenarios and specific applications, they inherently suffer from limited generalization ability, particularly when faced with sophisticated joint-modality deepfakes that maintain high internal consistency between video and audio. One of the major limitations of unimodal strategies is their inability to capture cross-modal correlations—critical cues that can arise when the manipulation involves both audio and video streams simultaneously [7].

In contrast, multimodal deepfake detection techniques leverage information from multiple sources, typically combining visual and auditory features to create a more comprehensive and resilient detection framework. These methods aim to overcome the limitations of unimodal models by exploiting the synergy between modalities to improve the model's robustness and accuracy. Recent studies have explored the use of cross-modal embedding spaces that can effectively represent and differentiate between authentic, audio-forged, and video-forged content. As illustrated in Fig. 1, such embedding spaces allow forged samples to be mapped along distinct vector directions, thereby enhancing the discriminability of various types of manipulations [8], [9]. Despite these promising advancements, current multimodal approaches often focus predominantly on the fusion of features from different modalities, without paying sufficient attention to the individual optimization and calibration of the modality-specific feature extractors. This oversight can lead to suboptimal feature representations, limiting the overall detection performance.

Furthermore, the limitations of existing facial forgery datasets introduce additional obstacles to the development of robust deepfake detection systems. A prevalent issue in many benchmark datasets is the limited diversity of manipulation techniques and the distribution mismatch between training and testing scenarios. Existing datasets often contain deepfakes generated by a fixed set of synthesis algorithms, which may not adequately represent the rapidly evolving landscape of deepfake generation methods encountered in real-world applications. For instance, models trained on datasets like KoDF [10] or DFDC [11] may perform well on known manipulation techniques but struggle significantly when confronted with novel synthesis algorithms that were not present during training. This generalization gap poses significant challenges for deployment in practical scenarios, where new deepfake generation methods are constantly emerging and the statistical distributions of manipulated content can differ substantially from those seen during training. Moreover, most existing detection models primarily rely on spatial features extracted from raw video frames, often neglecting valuable cues derived from UV texture maps. UV maps, which represent the geometric unwrapping of a 3D facial surface, contain rich and fine-grained texture details that are especially useful for capturing subtle inconsistencies introduced during forgery. Integrating UV texture information into the feature extraction pipeline holds the potential to significantly enhance the precision and reliability of deepfake detection systems, particularly when dealing with sophisticated synthesis methods that may evade detection based solely on spatial features.

To address these problems, we propose a novel framework, FUME. This framework employs TAViT to extract spatial and temporal features from videos, incorporating facial UV maps to capture subtle forgery traces in visual content. Moreover, it uses AST [12] to extract fine-grained spectral characteristics from audio, enhancing the detection of audio forgeries. The features from both modalities are aligned through contrastive learning, further refined via a cross-modal classifier for deep integration. Moreover, a momentum-based distillation mechanism is introduced to significantly enhance the robustness and generalization performance of the model. our main contributions are summarized as follows:

1) We propose a multimodal fusion framework leveraging facial UV maps generated by the 3DDFA model [13]. By integrating these textures with video frames via shared-weight spatial encoders, our approach captures fine-grained spatial-temporal features.
2) A novel video-audio contrastive learning framework is introduced to align embeddings and enhance cross-modal representation, incorporating a momentum model for consistency.
3) We employ a One-Class Softmax loss that offers significant advantages over traditional binary classification for deepfake detection. This approach learns a compact hypersphere representation of authentic video features in the embedding space, enabling detection through deviation measurement rather than boundary classification.
4) Extensive experiments, including ablation studies and comparisons against state-of-the-art methods, validate the robustness and superior performance of our approach across benchmarks.

II. Related Work

A. Deepfake Detection

Recent developments in deepfake detection have witnessed significant strides in single-modal approaches, with researchers dedicating substantial efforts to visual and audio-based detection strategies. In the visual domain, convolutional neural networks (CNNs) remain the cornerstone of many deepfake detection architectures. For instance, the integration of CNNs with reinforcement learning frameworks such as Deep Q-Networks (DQNs) has shown promising results, particularly in scenarios involving static image manipulation detection, where the model can learn optimal policies to identify forgery cues through iterative exploration [14]. Other studies have combined classical backbone networks like ResNet50 with person-of-interest (POI) techniques to enhance identity-specific detection performance, thereby allowing the system to focus on facial features that are more likely to reveal inconsistencies in generated content [15]. More recent innovations involve the hybridization of CNNs with Transformer-based architectures, notably the Vision Transformer (ViT), which excels in capturing long-range dependencies and global representations. This fusion enables detectors to leverage both local pixel-level textures and global semantic features, as evidenced in works integrating CNN and ViT [16], as well as those combining EfficientNet-B7 with ViT to balance detection performance and computational efficiency [17]. Beyond traditional classification pipelines, techniques such as depth map-guided triplet networks [18] and artifact-disentangled adversarial learning [19] aim to explicitly learn the spatial characteristics of forgery artifacts, particularly in complex scenarios involving face-swapping or identity morphing.

In parallel, audio-based deepfake detection has emerged as a critical research area, especially given the proliferation of synthetic voices capable of mimicking human speakers with alarming realism. Yi et al. [20] proposed a comprehensive taxonomy that classifies audio deepfakes into five distinct categories: Text-to-Speech (TTS), Voice Conversion (VC), Emotion Fake, Scene Fake, and Partially Fake. Each category presents unique challenges for detection, stemming from differences in synthesis methods, prosodic features, and contextual coherence. Early detection efforts focused primarily on handcrafted acoustic features such as Mel-Frequency Cepstral Coefficients (MFCC), Linear Frequency Cepstral Coefficients (LFCC), and Constant-Q Cepstral Coefficients (CQCC) [21], which capture short-term spectral properties of speech. However, as manipulation techniques evolved, these traditional features proved insufficient for generalized detection. Researchers then incorporated long-term spectral features [22], prosodic characteristics [23], and deep neural embeddings [24] to enrich the detection signal. With the rise of deep learning, CNN-based architectures [25], ResNet variants [26], and graph neural networks such as AASIST [27] became popular due to their superior representation capacity. End-to-end systems like RawNet2 [28] and AASIST [27] jointly optimize feature learning and classification, achieving competitive performance across several public benchmarks. However, the generalization of these models to unseen spoofing attacks and variable acoustic environments remains a persistent challenge. To address this, recent studies have proposed domain-aware loss functions [29] and continual learning mechanisms [30] to improve model robustness and adaptability across diverse conditions.

While single-modal deepfake detection approaches have undeniably laid the foundation for understanding and mitigating visual and auditory forgeries, their inherent design—focusing exclusively on a single modality—renders them insufficient for coping with increasingly sophisticated and synchronized cross-modal manipulations. Modern deepfakes often exhibit coherent coordination between facial expressions, lip movements, and speech content, making it challenging for single-modal systems to identify discrepancies that are subtle yet semantically inconsistent. This limitation is primarily due to the absence of inter-modal correlation modeling in unimodal systems, which prevents them from capturing mismatches between the temporal dynamics of visual cues and the phonetic or prosodic elements of audio streams. As deepfake generation techniques advance toward seamless integration across modalities, the need for more holistic detection frameworks becomes imperative. Consequently, researchers have turned their attention toward multimodal deepfake detection strategies, which are inherently more capable of leveraging cross-modal dependencies to improve detection accuracy and robustness in real-world scenarios.

In contrast to unimodal systems, multimodal deepfake detection methods integrate heterogeneous sources of information—such as audio, visual, and spatio-temporal signals—into a unified detection framework. These methods aim to exploit the synergy between modalities to better detect manipulations that may not be apparent when modalities are considered in isolation. For instance, AVoiD-DF [31] combines auditory and visual features to uncover inter-modal inconsistencies that signal manipulation, while TruFor [32] employs a multi-clue fusion mechanism that not only enhances detection but also enables the localization of tampered regions within frames. More advanced techniques such as dynamic graph learning with frequency relation reasoning [33] focus on learning temporal dependencies between video frames, providing fine-grained detection of temporal discontinuities. Similarly, HolisticDFD [2] leverages spatial-temporal transformers to simultaneously model intra-frame artifacts and inter-frame dynamics, thereby enhancing its ability to detect both static and temporal forgeries. Other works like SI-Net [9] emphasize spatial interactions across modalities, demonstrating effectiveness in scenarios with complex video manipulation. Collectively, these multimodal approaches present a significant step forward in combating high-quality deepfakes that evade unimodal detectors. However, despite their promise, many existing methods still face challenges in generalizing to unseen domains, attack types, and environmental conditions. Current research has yet to fully explore how to optimize cross-modal feature alignment and adapt detection frameworks to maintain robustness under varied real-world conditions. To solve these problems, the FUME framework proposed in this paper offers advantages by jointly analyzing audio and visual streams, utilizing facial UV maps via TAViT for enhanced visual detail detection, and employing cross-modal alignment to capture discrepancies missed by single-modality methods.

B. Multimodal Representation Learning

Multimodal representation learning, particularly involving audio-visual content, has become increasingly vital for deepfake detection as manipulated media grows in sophistication. Manzoor et al. [34] surveyed the evolution of multimodal approaches, highlighting that the concurrency between modalities provides critical self-supervision signals for deepfake detection. Several pioneering works have leveraged this insight, with Zhou et al. [35] demonstrating that inconsistencies between lip movements and speech can be powerful indicators of manipulation. Li et al. [36] extended this by proposing a temporal sync network that captures audio-visual synchronization patterns to detect deepfakes. However, as Qian et al. [37] noted, directly transferring information between modalities may lead to conflicts and redundancy, causing false positives in detection systems. Audiovisual speech recognition research by Shi et al. [38] introduced AV-HuBERT, employing hybrid architectures to extract joint features from audio and visual streams, while Chung et al. [39] developed complementary techniques for cross-modal alignment specific to speech manipulation detection.

Contrastive learning has emerged as a powerful paradigm for multimodal deepfake detection, with Jaiswal et al. [40] implementing contrastive objectives to identify unnatural co-occurrences in manipulated content. This approach was further refined by Zhuge et al. [41] through CAV-MAE, which explicitly leverages audio-visual pair information to detect inconsistencies. Cross-modal attention mechanisms, pioneered by Mittal et al. [42] and refined in AV-transformer by Lin et al. [43], established stronger correlations between modalities to detect deepfakes. More sophisticated fusion strategies have been explored by Wang et al. [44] who proposed graph-based approaches similar to VQA-GNN for modeling complex relationships between audio and visual elements. Appearance-

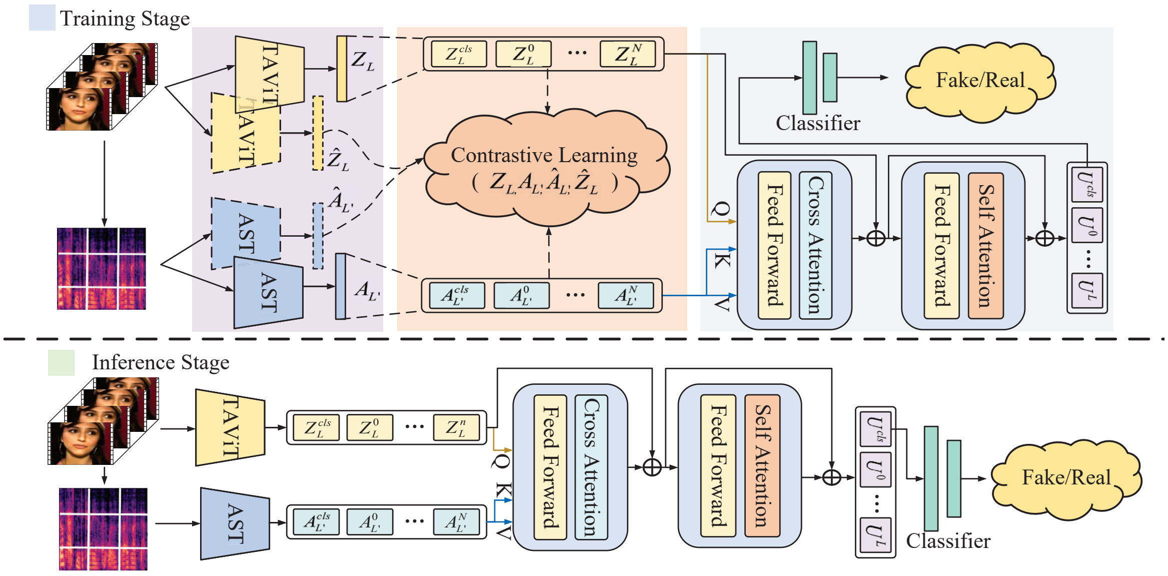

Fig. 2. Overview of the proposed FUME Framework. It **1)** extracts spatial-temporal features from facial images and UV maps via the Vision Encoder using TAViT, and **2)** processes audio signals using the AST-based Audio Encoder to capture frequency and temporal information. These features are fused through cross-attention in the multimodal Fusion Classifier for robust forgery detection. The training stage employs a contrastive learning framework to align visual and audio representations, enhancing modality consistency. Modules with dashed lines indicate the momentum versions of TAViT and AST, respectively.

based methods by Zhao et al. [45] localize audio-visual inconsistencies through correlation of embeddings, while Zhou et al. [46] developed temporal consistency verification through transformer architectures like VATT. Haliassos et al. [47] introduced a novel lip-sync error detection system specifically for deepfake videos. Kumar et al. [48] demonstrated that cross-modal consistency verification significantly outperforms single-modality approaches, while Dolhansky et al. [49] incorporated multimodal transformers to detect manipulated facial expressions and corresponding audio. As media manipulation techniques advance, Agarwal et al. [50] and Chen et al. [51] suggest that future deepfake detection must focus on more inherent relationships between audio-visual modalities to improve detection robustness. While existing multimodal methods have advanced the field, FUME further enhances performance by incorporating geometrically rich facial UV maps before fusion and employing a more stable momentum contrastive learning for precise cross-modal semantic alignment, leading to improved robustness against sophisticated manipulations.

III. Method

Fig.2 illustrates the overall architecture of the proposed FUME network. It consists of a visual encoder built on the TAViT, an audio encoder leveraging the AST, and a multimodal fusion classifier. These components collaboratively exploit visual and audio modalities to enhance forgery detection by capturing complementary features. The functional architectures of each constituent component are delineated as follows:

- The proposed TAViT leverages 3D facial maps to simultaneously extract spatial and temporal features from video frames. By exploiting the rich geometric information embedded in 3D facial maps, the module comprehensively captures the intricate spatiotemporal representations of facial dynamics. The meticulously embedded features serve as the foundational resource for subsequent multimodal fusion, enabling a nuanced representation of visual characteristics.
- The AST module transforms raw audio signals into spectrographic representations, subsequently employing a Vision Transformer (ViT) architecture for feature extraction. It translates the temporal audio signal into a spatial-frequency domain, allowing the transformer to capture complex spectral patterns and contextual audio representations. The transformation facilitates a sophisticated feature learning process that transcends traditional audio feature extraction methodologies.
- The multimodal classifier orchestrates the critical task of heterogeneous feature alignment and fusion. It initially applies a novel contrastive learning framework specifically designed for deepfake detection, which systematically aligns the disparate features extracted by TAViT and AST. Subsequently, a cross-attention mechanism is deployed to achieve deep feature fusion, enabling complex inter-modal interactions. This sophisticated approach culminates in a robust multimodal deepfake identification process, leveraging the complementary information from visual and audio domains.

A. Preprocessing

For the preparation of our audio-visual data inputs, we implemented a multi-stage preprocessing pipeline designed to isolate pertinent information and enhance model robustness. Initially, visual frames and corresponding audio waveforms were synchronously extracted from the raw video sources at sampling rates of 5 frames per second (fps) and 16 kHz, respectively. The 5 fps rate for the visual stream was chosen to capture significant facial dynamics relevant to speech while maintaining computational tractability, whereas the 16 kHz audio sampling rate ensures sufficient coverage of the frequency spectrum characteristic of human speech for subsequent acoustic feature extraction.

Subsequently, the visual frames underwent processing to accurately localize and isolate the facial region, thereby mitigating interference from irrelevant background elements which typically exhibit minimal correspondence with the spoken audio content. To achieve robust face detection across a wide range of scales, particularly challenging small faces often encountered in video data, we employed the Single Shot Scale-invariant Face Detector (S^{3}FD) [52]. S^{3}FD's effectiveness stems from its scale-equitable architecture, which utilizes anchors across multiple convolutional layers with varying strides (4 to 128 pixels). Its anchor design is informed by principles of effective receptive field matching and equal-proportion intervals, ensuring adequate feature support for faces of diverse sizes. Furthermore, S^{3}FD incorporates a scale compensation anchor matching strategy to improve recall rates for faces whose sizes fall between predefined anchor scales, and utilizes a max-out background label specifically for the anchors associated with detecting the smallest faces, significantly reducing the false positive rate often caused by the necessary density of small anchors. Following detection, the identified facial regions were cropped from the frames, yielding sequences of face-centric images.

To further enhance the model's generalization capability and resilience to natural variations and unforeseen corruptions common in real-world visual data, we applied data augmentation to the cropped facial images using the AugMix technique [53]. AugMix operates by generating multiple and diverse augmentation sequences, each composed of a stochastic chain of basic augmentation operations and mixing their outputs. Specifically, several augmented versions of an image are created via different augmentation chains. These versions are then combined using a convex combination with weights sampled from a Dirichlet distribution. This composited image is subsequently interpolated with the original cropped face image, using a mixing weight sampled from a Beta distribution. This strategy produces augmented samples that exhibit significant diversity, yet remain close to the original data manifold, effectively promoting model robustness without sacrificing representation fidelity.

During visual processing, the extracted 16 kHz audio waveforms were transformed into a suitable time-frequency representation. We computed log-mel spectrograms, a standard and perceptually motivated acoustic feature. This process involves applying a Short-Time Fourier Transform (STFT) to the waveform, mapping the resulting spectral power onto the Mel scale using a filter bank to mimic human auditory perception, and finally taking the logarithm to compress the dynamic range and approximate loudness perception. This resulted in spectrograms with several frequency bins, effectively capturing the temporal evolution of the spectral energy distribution crucial for characterizing speech. Henceforth, the preprocessed visual sequences, consisting of S^{3}FD-cropped and AugMix-augmented facial frames, and the corresponding log-mel spectrograms are referred to as the visual and audio input representations, respectively, serving as the direct input to our audio-visual learning model.

B. Visual Encoder using TAViT

Traditional deepfake detection methodologies often operate solely on facial image data extracted from individual video frames. While effective to some extent, these approaches may overlook subtle yet crucial geometric inconsistencies and texture artifacts that are characteristic of sophisticated forgery techniques. Identifying high-quality deepfakes necessitates a finer-grained analysis, as manipulations might manifest as minute discrepancies in facial contours during expressions, unnatural surface textures, or inconsistencies in lighting across the 3D facial geometry. These subtle cues are often difficult to capture from 2D image representations alone.

To overcome these limitations, we propose the Texture-Aware Video Transformer (TAViT), a dedicated visual encoder module illustrated in Fig. 3. TAViT is designed to extract and integrate rich spatio-temporal visual features by concurrently processing standard facial images and corresponding facial UV maps. Facial UV maps serve as a 2D representation of the unwrapped 3D facial surface texture and geometry. Their inclusion provides enhanced spatial granularity, enabling the model to scrutinize fine details related to facial expressions, surface texture variations, and lighting interactions across the facial geometry, which are often critical for detecting advanced forgeries. the facial UV maps are generated frame-wise using the 3DDFA from corresponding facial images.

The processing pipeline for each frame t of the input video utilizes the input facial image and its corresponding facial UV map. These two distinct data modalities, capturing appearance and detailed surface texture respectively, are then prepared for transformer-based analysis. Each modality is independently subjected to the standard Vision Transformer (ViT) tokenization process: division into a fixed grid of non-overlapping patches followed by linear projection to form patch embeddings. The sequence of embeddings derived from the facial image and the sequence derived from the UV map are subsequently processed by two parallel Vision Transformer (ViT) encoder branches. Critically, these two branches operate with shared weights. This weight-sharing strategy serves multiple purposes: it promotes parameter efficiency, enforces a consistent feature extraction paradigm across the visual appearance and the geometric/textural data, and facilitates the learning of correlated features, ultimately ensuring that the resulting representations from both branches are directly comparable and suitable for effective downstream fusion.

To be specific, the construction of the input sequence for each Vision Transformer (ViT) branch at a specific frame t is a critical first step. This sequence, denoted as $F_{0,t}^{\text{mod}}$, serves as the input to the first Transformer encoder layer ($\ell = 0$). It commences with a special, learnable embedding vector known as the class token, F_{cls}. This token is prepended to the sequence and is designed to aggregate global information about the entire input (either the facial image or the UV map) as it passes through the subsequent layers. It aggregates global facial information and enables cross-frame temporal modeling. It captures discriminative regions that indicate potential manipulation. Following the class token are the patch embeddings $f_{i,t}^{\text{mod}}$, where each $f_{i,t}^{\text{mod}}$ is the result of linearly projecting the i-th flattened image patch of the respective modality (mod $\in \{\text{face}, \text{UV}\}$). To encode the crucial spatial information regarding the original position of each patch, which is otherwise lost in the sequential representation, learnable positional embeddings, collectively denoted by E_p but specifically applied per position ($E_{p,i}$ for the i-th token including $E_{p,0}$ for F_{cls}), are added element-wise to the corresponding patch embeddings (and the class token). For each patch at position i in the flattened sequence, we add a learnable positional embedding $E_{p,i}$. This encoding preserves spatial relationships crucial for detecting geometric inconsistencies. The input formulation is therefore:

$$F_{0,t}^{\text{mod}} = \left[F_{\text{cls}}; f_{1,t}^{\text{mod}} + E_{p,1}; f_{2,t}^{\text{mod}} + E_{p,2}; \cdots; f_{T,t}^{\text{mod}} + E_{p,t}\right], \tag{1}$$

where P represents the total number of patches extracted from the input modality.

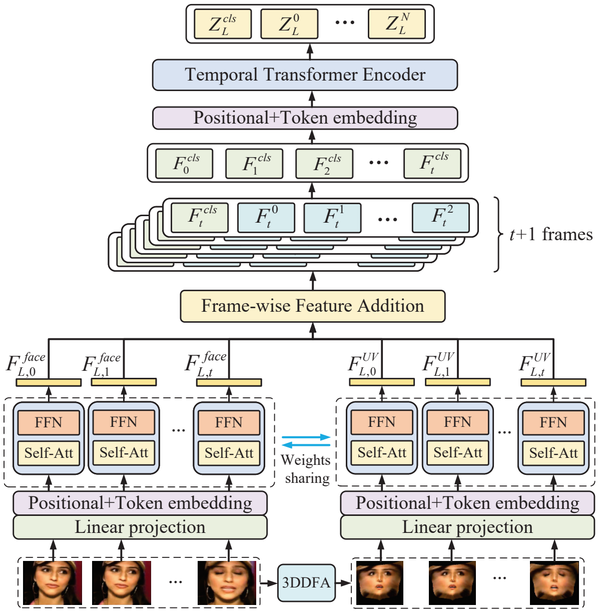

Fig. 3. The Texture-Aware Video Transformer (TAViT). This module extracts and integrates spatial-temporal features from video frames and facial UV maps. The spatial features are first processed by two Vision Transformer branches with shared weights. Features from the facial image and UV map branches are fused via frame-wise feature addition, producing a unified representation. Temporal dependencies across frames are then captured using a Temporal Transformer Encoder. The output embeddings encapsulate both spatial and temporal information for subsequent multimodal fusion.

Both input sequences $F_{0,t}^{\text{face}}$ and $F_{0,t}^{\text{UV}}$ undergo processing through a stack of L Transformer encoder layers with shared weights between the parallel branches. This design choice ensures consistent feature extraction paradigms across visual appearance and geometric texture data. The multi-layer architecture enables hierarchical feature refinement, where each successive layer captures more sophisticated spatial relationships and subtle manipulation artifacts.

$$F_{\ell,t}^{\text{mod}} = \text{MSA}(\text{LN}(F_{\ell-1,t}^{\text{mod}})) + F_{\ell-1,t}^{\text{mod}}, \quad \ell = 1, \ldots, L, \tag{2}$$

where $F_{\ell,t}^{\text{mod}}$ represents the feature embeddings at layer ℓ for either the facial image or facial UV maps at frame t.

To obtain a unified representation, the spatial features from the facial image and facial UV maps branches are fused via element-wise addition. This results in the fused feature map:

$$F_t = F_{L,t}^{\text{face}} + F_{L,t}^{\text{UV}}, \tag{3}$$

where F_t encapsulates the complementary strengths of both modalities, combining the detailed spatial characteristics captured by facial images with the fine-grained geometric and texture information provided by facial UV maps.

After fusing the spatial features from both the facial image and facial UV maps branches, the class tokens from each frame are extracted and concatenated to form the input to the Temporal Encoder. Let F_t^{cls} represent the class token from the fused feature F_t for frame t. Mathematically, this can be expressed as:

$$Z_0 = [F_1^{\text{cls}}; F_2^{\text{cls}}; \cdots; F_N^{\text{cls}}], \tag{4}$$

where Z_0 represents the class tokens from all N frames at layer L. These class tokens are then passed through the Temporal Encoder, which can be formulated as follows:

$$Z_\ell = \text{MSA}(\text{LN}(Z_{\ell-1})) + Z_{\ell-1}, \quad \ell = 1, \ldots, L, \tag{5}$$

where Z_ℓ represents the temporal features at layer ℓ. The outputs Z_L encapsulate both temporal and spatial information derived from visual features, which are further utilized for cross-modal joint learning.

The final output sequence Z_L from the Temporal Transformer Encoder represents the deeply integrated spatio-temporal visual features. Each element in Z_L corresponds to a frame but is contextually aware of the entire video sequence's visual information, derived from both facial images and UV maps. These powerful visual embeddings Z_L are then utilized for subsequent multimodal fusion or downstream classification tasks, providing a robust foundation for detecting sophisticated video forgeries.

C. Audio Encoder based on AST

Traditional audio feature extraction techniques predominantly rely on convolutional neural networks (CNNs), which, despite their effectiveness in capturing local features, exhibit inherent limitations in modeling long-range dependencies and global contextual information within audio signals. CNNs operate with fixed receptive fields, which restrict their ability to encode relationships between distant time frames—a crucial aspect in understanding complex auditory patterns such as spoken language, environmental soundscapes, or musical sequences.

To overcome these limitations, we adopt the Audio Spectrogram Transformer (AST), a transformer-based architecture specifically designed for processing audio spectrograms. Inspired by the Vision Transformer (ViT) and adapted to the audio domain, AST leverages the self-attention mechanism to model global dependencies and learn more holistic audio representations. This is particularly valuable in capturing intricate temporal and frequency structures in auditory data.

In our approach, raw audio waveforms are first transformed into Mel-frequency filter bank (fbank) features using standard signal processing techniques. These fbank features constitute a two-dimensional representation of the signal, formulated as a matrix $A \in \mathbb{R}^{T \times M}$, where T denotes the number of time frames and M indicates the number of frequency bins. Each row in this matrix captures the frequency content at a specific time slice, resulting in a time-frequency representation suitable for transformer-based modeling.

To make the fbank features compatible with the transformer encoder, we linearly project each time-frequency vector into a fixed-dimensional embedding space. A special learnable class token A_{cls}, analogous to the [CLS] token used in natural language processing transformers, is prepended to the sequence of embeddings. This token is designed to aggregate global information across all time frames during self-attention operations and is used for downstream prediction tasks.

Additionally, positional embeddings E_p are incorporated into the input sequence to retain information about the temporal order of frames, which is otherwise lost due to the permutation-invariant nature of self-attention. The initial input to the transformer encoder can be expressed as:

$$A_0 = [A_{\text{cls}} + E_p^{(0)}; a_1 + E_p^{(1)}; \dots ; a_T + E_p^{(T)}], \tag{6}$$

where a_i denotes the embedding of the i-th fbank frame, and $E_p^{(i)}$ is the positional embedding for the corresponding position.

The resulting sequence of embeddings is passed through a stack of L' transformer encoder layers, each consisting of a multi-head self-attention (MSA) mechanism followed by a position-wise feed-forward network. Layer normalization (LN) and residual connections are applied at each stage to facilitate training stability and information flow. The transformation at each layer ℓ can be formalized as:

$$A_\ell = \text{MSA}(\text{LN}(A_{\ell-1})) + A_{\ell-1}, \quad \ell = 1, \dots, L'. \tag{7}$$

After processing through the full stack of transformer layers, the final output sequence $A_{L'}$, particularly the representation corresponding to A_{cls}, is treated as a compact summary of the audio signal. This output can be directly utilized or integrated with visual modality representations, such as Z_L, in a cross-modal fusion framework for multimodal learning tasks.

D. multimodal Fusion Classifier

In the multimodal fusion process, the video embeddings Z_L and audio embeddings $A_{L'}$, are used to compute cross-attention. The attention mechanism is applied to the normalized query (Q), key (K), and value (V) features, and is formulated as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{K^T Q}{\sqrt{D}}\right) V, \tag{8}$$

where we perform cross-attention between video embeddings Z_L and audio embeddings A_L by designating the video embeddings as the query (Q) and the audio embeddings as the key (K) and value (V), as described below:

$$U = \text{Attention}(Z_L, A_{L'}, A_{L'}) + Z_L. \tag{9}$$

Finally, the output obtained after the cross-attention mechanism is passed through the self-attention mechanism and feed-forward sublayers. The `[CLS]` token from the resulting embedding is then extracted and fed into the classifier to produce the final classification result.

E. Loss Function

To address the challenge of insufficient and inconsistent negative samples in contrastive learning and ensure the stability and precision of the learned representations, We adopted and enhanced a multimodal momentum-driven contrastive learning framework introduced by Shao et al. [54], which helps maintain stable representations across training iterations, enhancing the alignment between audio and video modalities. Contrastive learning aims to bring together positive pairs and push apart negative pairs. Importantly, in the case of manipulated videos, the audio-video pairs are also treated as negative samples, as the audio and video no longer correspond coherently.

The contrastive loss is based on the InfoNCE loss. For example, the audio-to-video loss is formulated as:

$$\begin{aligned} &L_{\text{a2v}}(A_{L'}, Z_L^+, Z_L^-) = \\ &\quad - \mathbb{E}_{p(A_{L'}, Z_L)} \left[\log \frac{\exp(S(A_{L'}, Z_L^+)/\tau)}{\sum_{k=1}^{K} \exp(S(A_{L'}, Z_L^-)/\tau)} \right], \end{aligned} \tag{10}$$

where the τ controls the sharpness of the similarity distribution, Z_L^+ denotes a positive video sample matched to the audio, while Z_L^- represents a set of negative samples, including randomly selected or manipulated videos. $S(A_{L'}, Z_L)$

represents the similarity score between the audio embedding and the video embedding, The semantic alignment between heterogeneous modalities is established through cross-modal projection mechanisms where the class token functions as the unified semantic anchor. Two parameterized mapping functions h_v and $\hat{h}_a$ project the visual and audio class tokens into a shared 256-dimensional latent space for similarity quantification. Therefore, $S(A_{L'}, Z_L)$ can be formulated as:

$$S(Z_L, A_{L'}) = h_v(Z_L^{cls})^\top \hat{h}_a(\hat{A}_{L'}^{cls}), \tag{11}$$

where $\hat{h}_a$ is class token from audio momentum encoders and $\hat{h}_a(\hat{A}_{L'}^{cls})$ represents projected audio embeddings from audio momentum projection head.

This architecture explicitly encodes modality-specific features into isomorphic vector representations, enabling efficient cross-domain comparison through temperature-scaled cosine similarity. The learnable projection parameters h_v and h_a are optimized through contrastive learning, establishing robust alignment between visual and auditory semantic spaces. Similarly, video-to-audio contrastive loss is as follows:

$$\begin{aligned} &L_{\text{v2a}}(Z_L, A_{L'}^+, A_{L'}^-) = \\ &\quad - \mathbb{E}_{p(Z_L, A_{L'})} \left[\log \frac{\exp(S(Z_L, A_{L'}^+)/\tau)}{\sum_{k=1}^{K} \exp(S(Z_L, A_{L'}^-)/\tau)} \right]. \end{aligned} \tag{12}$$

Finally, we incorporate the losses for other pairings to form ManipulationAware Contrastive Loss: L_{v2a} for video-to-audio, L_{v2v} for video-to-video, and L_{a2a} for audio-to-audio. The total contrastive loss is the average of these four terms:

$$L_{mac} = \frac{1}{4} \left(L_{a2v} + L_{v2a} + L_{v2v} + L_{a2a} \right). \tag{13}$$

To address the challenge of detecting unseen deepfake attacks, we adopt a One-Class Softmax loss, building on the work of Zhang et al. for speech deepfake detection [55]. The key motivation is that traditional binary classification approaches may overfit to specific manipulation techniques present in the training data, leading to poor generalization when confronted with novel deepfake generation algorithms in real-world scenarios. The OC-Softmax loss is specifically designed to learn a feature embedding space where real samples are compactly clustered around a learned center, while manipulated samples are pushed away from this cluster by an angular margin. Unlike conventional binary classification losses that attempt to learn separate decision boundaries for both real and fake classes, OC-Softmax focuses on characterizing the real samples distribution and treats any significant deviation from this compact representation as potential manipulation. This approach is particularly well-suited for deepfake detection because the distribution of authentic data is relatively stable and well-characterized, whereas the space of possible manipulations is vast and continuously evolving with new synthesis techniques. The classification loss is defined as:

$$L_{ocs} = \frac{1}{N} \sum_{i=1}^{N} \log \left(1 + e^{\alpha(m_{y_i} - \hat{w}^T \mathbf{x}_i)(-1) y_i} \right), \tag{14}$$

where $\mathbf{x}_i$ is the normalized weight vector for the target class, and $\hat{w}$ is the weight vector for the target class. The margin m_{y_i} increases the separation between real and manipulated samples, while α is a scaling factor.

To enhance the quality of representation in each modality before fusion, we implement a Modality-Specific Classification Loss (L_{msc}) that directly supervises each unimodal encoder:

$$\begin{aligned} L_{msc} = &\frac{1}{S^v} \sum_{k=1}^{S^v} \text{CE} \left(F_V(cls_k^v), y_k^v \right) \\ &+ \frac{1}{S^a} \sum_{k=1}^{S^a} \text{CE} \left(F_A(cls_k^a), y_k^a \right), \end{aligned} \tag{15}$$

where CE denotes the cross-entropy function, cls_k^v represents the k-th class token Z_L^{cls} from the final layer of the TAViT, cls_k^a represents the class token $A_{L'}^{cls}$ from the final layer of the AST. The functions F_V and F_A are modality-specific classification heads that map the class token representations to class probability distributions. Here, S is the number of tokens in the training set and y represents the ground truth of the modalities.

This architectural design leverages the natural summarization capability of transformer class tokens, which aggregate contextual information across their respective modality sequences. By applying classification objectives directly to these tokens before fusion, we encourage each encoder to develop robust discriminative capabilities independently. This approach significantly improves the model's resilience to manipulations that may disproportionately affect one modality over another.

The total loss is a weighted sum of the momentum-based contrastive loss L_{mac}, the joint audio-visual classification loss L_{ocs} and the modality specific classification loss, defined as:

$$L_{total} = \alpha L_{mac} + \beta L_{ocs} + \gamma L_{msc}, \tag{16}$$

where L_{total} is the final cross-modal loss function for manipulation detection, and the parameters α, β and γ are trainable weights designed to balance the contributions of different loss functions.

IV. Experiment

A. Dataset and Implementation

We use three key datasets for deepfake detection: DeepfakeTIMIT [64], DFDC [11] and KoDF [10]. The DeepfakeTIMIT dataset contains both low-quality (64×64) and high-quality (128×128) face-swapped videos, with a total of 640 videos generated from the VidTIMIT database. The original audio tracks are retained for synchronized audio-visual analysis. The DFDC dataset (DeepFake Detection Challenge) comprises over 100,000 videos created using various deepfake and GAN-based methods, involving over 3,000 actors in diverse scenarios, providing a comprehensive benchmark for deepfake detection. The KoDF dataset (Korean DeepFake Detection Dataset) comprises 62,166 real videos and 175,776 fake videos

TABLE I

OVERALL COMPARISON WITH STATE-OF-THE-ART METHODS. THE METRICS ARE ACC AND AUC SCORES (%).

Method	Modality	DF-TIMIT (LQ)		DF-TIMIT (HQ)		DFDC		KoDF	
		ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Capsule [56]	$\mathcal{V}$	48.25	47.99	50.25	50.98	74.21	75.62	90.28	91.36
Xception [57]	$\mathcal{V}$	97.96	98.10	95.20	95.60	80.54	79.34	92.36	93.19
CViT [58]	$\mathcal{V}$	97.20	98.25	98.01	98.73	62.14	63.17	—	—
FTCN [59]	$\mathcal{V}$	85.45	87.23	90.56	90.69	73.18	74.03	77.49	78.11
Face X-ray [60]	$\mathcal{V}$	—	96.95	—	94.47	43.42	59.36	—	—
LipForensics [47]	$\mathcal{V}$	92.75	93.03	**99.25**	99.27	71.32	73.56	93.16	93.72
RealForensics [61]	$\mathcal{V}$	97.65	97.95	98.36	98.47	80.47	79.24	94.17	94.67
EmoForen [62]	$\mathcal{AV}$	92.76	92.35	94.79	95.46	82.33	84.40	93.26	94.41
MDS [63]	$\mathcal{AV}$	93.14	93.06	94.23	94.86	87.80	86.52	95.68	95.24
AVA-CL [7]	$\mathcal{AV}$	97.79	98.99	96.53	**99.86**	84.20	88.64	94.78	88.64
AVoiD-DF [31]	$\mathcal{AV}$	97.31	99.07	97.43	99.44	83.20	86.64	95.38	96.61
MCL [4]	$\mathcal{AV}$	97.76	97.20	98.91	99.09	86.45	86.61	96.78	97.64
Ours	$\mathcal{AV}$	**98.54**	**99.08**	98.86	**99.87**	**89.13**	**90.21**	**97.59**	**98.23**

generated using six different synthesis methods, featuring 403 subjects, providing a comprehensive resource for deepfake detection research.

The datasets are split into training, validation, and test sets with a 8:1:1 ratio, ensuring balanced real and fake samples in the test set. All experiments are conducted under consistent conditions for comparability.

B. Comparative Experiments

To evaluate the performance of our proposed method, we conducted comparative experiments with other unimodal and multimodal deepfake detection approaches on three benchmarks and the results are presented in Table I.

Comparison with Visual-only Methods. Our method significantly outperforms all visual-only approaches across all datasets. Capsule [56] utilizes capsule networks but demonstrates poor generalization, achieving only 48.25% accuracy on DF-TIMIT (LQ). Xception [57] employs domain-specific knowledge, achieving 97.96% accuracy on DF-TIMIT (LQ) but deteriorating to 80.54% on DFDC. CViT [58] uses convolutional vision transformers, obtaining 98.01% accuracy on DF-TIMIT (HQ) but substantially decreasing to 62.14% on DFDC. FTCN [59] leverages transformer architecture to capture long-range temporal dependencies, achieving 85.45% accuracy on DF-TIMIT (LQ), 73.18% on DFDC. Face X-ray [60] employs X-ray detection to reveal face swapping traces, achieving AUC scores of 96.95% and 94.47% on DF-TIMIT datasets but only 59.36% on DFDC. LipForensics [47] targets semantic irregularities in mouth movements, achieving 99.25% accuracy on DF-TIMIT (HQ) but dropping to 71.32% on DFDC and 93.16% on KoDF. RealForensics [61] proposes a hybrid approach using multi-modal pre-training with audio and visual data from real samples to learn internal representations, achieving 98.36% accuracy on DF-TIMIT (HQ), and 94.17% accuracy on KoDF. In contrast, our method achieves 98.54% accuracy on DF-TIMIT (LQ), 89.13% accuracy on DFDC, and 97.59% accuracy on KoDF, demonstrating superior generalization capability across all datasets.

Comparison with Audio-visual Methods. Our method outperforms existing multimodal approaches, including recent state-of-the-art methods. EmoForen [62] extracts emotional features from both modalities but may misclassify authentic videos due to emotional variations. MDS [63] computes audio-visual dissimilarity but overlooks potential audio forgery. AVA-CL [7] represents a contrastive learning strategy that achieves comparable performance to ours on DF-TIMIT (HQ) with 99.86% AUC but falls short on DFDC and KoDF. AVoiD-DF [31] employs audio-visual joint learning with temporal-spatial encoders, achieving 97.31% accuracy on DF-TIMIT (LQ) and 95.38% accuracy on KoDF. MCL [4] utilizes multi-modal contrastive learning to reduce cross-modal gaps, achieving 97.76% accuracy on DF-TIMIT (LQ) and 96.78% accuracy on KoDF. Our method achieves superior performance with 98.54% accuracy on DF-TIMIT (LQ) and 97.59% accuracy on KoDF, outperforming all comparative methods. Overall, our proposed multimodal method achieves optimal performance across all benchmark datasets, with particularly strong results on the challenging DFDC dataset (89.13% accuracy, 90.21% AUC), validating its effectiveness and robustness.

C. Cross-Dataset Generalization

The generalization ability of deepfake detection methods is a critical metric. Current approaches, designed for specific artifacts, often fail against unseen forgeries, limiting their real-world applicability. Enhancing cross-method generalization remains a key challenge. To evaluate the generalization ability of our proposed model, we conduct cross-dataset evaluations, training the model on two datasets and testing it on an unseen one to ensure the test set contains forgery types and

TABLE II
COMPARISON OF AUC (%) AND ACC (%) METRICS ACROSS DATASETS: THE MODEL IS TRAINED ON TWO DATASETS AND EVALUATED ON A SEPARATE UNSEEN DATASET.

Method	DF-TIMIT		DFDC		KoDF	
	ACC	AUC	ACC	AUC	ACC	AUC
LipForensics [61]	83.85	84.61	71.26	74.30	75.03	74.70
RealForensics [47]	85.45	86.78	74.09	74.31	77.03	76.70
EmoForen [61]	83.68	84.40	71.14	72.44	78.35	79.21
MDS [63]	86.22	87.06	73.29	72.30	78.51	79.04
AVA-CL [7]	89.24	90.16	78.34	77.98	83.62	84.23
AVoiD-DF [65]	88.67	89.33	84.23	**86.34**	80.67	81.23
MCL [66]	90.24	red89.01	84.34	85.98	84.11	83.89
ours w/o L_{ocs}	88.13	87.45	82.63	83.12	82.98	81.18
ours	**91.27**	**91.45**	**85.63**	86.08	**85.43**	**86.61**

data distributions different from the training data. This setup simulates real-world scenarios with varying data sources.

Table II presents the AUC and ACC results of our model across different datasets. For clarity and focused comparison, we select representative visual-only methods that achieved the highest performance in our intra-dataset experiments: LipForensics [47] and RealForensics [61], while maintaining comprehensive comparisons with all audio-visual approaches. Our model demonstrates superior generalization performance across all benchmarks. On DF-TIMIT, our model achieves an AUC of 91.45% and an ACC of 91.27%, significantly outperforming competitive methods such as AVoiD-DF (89.33% AUC, 88.67% ACC) and MCL (89.01% AUC, 90.24% ACC). On the challenging DFDC dataset, our approach attains an AUC of 86.08% and an ACC of 85.63%, substantially surpassing existing methods including AVA-CL (77.98% AUC, 78.34% ACC) and MDS (72.30% AUC, 73.29% ACC). On KoDF, our model achieves robust cross-dataset performance with an AUC of 86.61% and an ACC of 85.43%, outperforming recent state-of-the-art methods like AVoiD-DF (81.23% AUC, 80.67% ACC) and MCL (83.89% AUC, 84.11% ACC).

The superior generalization ability stems from our one-class learning strategy, which addresses the distribution mismatch problem in deepfake detection. Unlike binary classification that assumes similar distributions for both real and fake classes, our approach focuses on learning compact representations for authentic data while pushing fake samples away. The ablation study "ours w/o L_{ocs}" confirms this effectiveness, showing performance drops to 81.18% AUC on KoDF when removing the L_{ocs}.

D. Ablation Study

To evaluate the contribution of each component within our proposed framework, we present an ablation study in Table III.

Comparison of Unimodal and Multimodal Approaches. Although some previous works have exploited only visual information for deepfake detection, they overlook the complex scenarios where both audio and visual modalities are manipulated simultaneously. In this paper, our method systematically reveals and leverages the inherent dependencies between audio and visual modalities. unimodal learning using only visual information achieves limited performance with an average of 76.76% ACC and 77.31% AUC across four datasets, while audio-only performance is even worse, reaching only 58.19% ACC and 58.32% AUC on average. In contrast, our proposed multimodal method achieves significant improvements on all datasets. This enhancement stems from our designed cross-modal attention mechanism, which effectively captures subtle patterns of audio-visual inconsistencies that are often overlooked in unimodal analysis. This demonstrates the critical value of audio-visual joint learning in building more robust deepfake detection systems.

Effectiveness of TAViT. The Texture-Aware Video Transformer (TAViT) represents one of our core innovations, specifically designed to capture spatiotemporal inconsistencies in video sequences. It first extracts spatial features from each frame. Then, the model extracts cross-frame temporal relationships through a temporal Transformer encoder. This design allows the model to first understand the spatial structure within each frame before analyzing patterns of change across the temporal dimension, thereby more precisely localizing forgery traces. Ablation experiments indicate that removing TAViT results in an average performance decrease of approximately 19.34% ACC and 19.69% AUC. This significant drop validates the exceptional capability of our TAViT architecture in capturing subtle temporal inconsistencies in deepfake videos, particularly for high-quality forgeries that are difficult to detect through static frame analysis. The success of TAViT lies in its ability to seamlessly integrate local spatial attention with global temporal modeling, which is crucial for identifying artifacts in modern deepfake techniques.

Contribution of Facial UV Mapping. Facial UV mapping is another innovation we introduced, providing a normalized facial texture representation that effectively addresses limitations of traditional methods when handling faces under varying poses and lighting conditions. UV mapping projects 3D facial geometry onto a standardized 2D space, making forgery traces more identifiable, especially in facial boundaries and texture transition regions. Removing the facial UV mapping component results in performance decreases of 4.01% ACC and 6.04% AUC on the DFDC dataset. These results demonstrate the unique value of facial UV mapping in capturing and amplifying subtle visual forgery traces in deepfake videos, which may be difficult to distinguish in the original pixel space. The advantage of facial UV mapping lies in its ability to provide pose-invariant consistent representations, enabling the model to focus on learning essential differences between real and forged faces, rather than being distracted by irrelevant factors such as pose and illumination variations.

Role of AST. The AST is a component specifically designed for processing the audio modality, based on the Transformer architecture to extract deep semantic features from audio signals. Compared to traditional convolutional neural networks,

TABLE III
ABLATION STUDY RESULTS ACROSS DIFFERENT DATASETS. THE METRICS ARE ACC AND AUC SCORES (%).

Method	DF-TIMIT (LQ)		DF-TIMIT (HQ)		DFDC		KoDF	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
visual only	77.89	78.97	78.35	79.58	73.51	74.32	77.28	76.36
audio only	57.43	57.88	57.43	57.88	58.54	60.34	59.36	57.19
w/o facial UV maps	97.43	96.22	95.34	95.23	85.14	84.17	95.54	96.88
w/o TAViT	79.45	78.23	77.35	80.29	71.42	70.56	75.16	78.65
w/o AST	94.93	94.95	95.67	95.12	84.55	84.96	92.08	92.46
w/o L_{mac}	94.76	95.35	95.79	96.46	85.33	84.40	94.26	93.91
w/o L_{ocs}	96.83	97.42	97.13	98.71	87.39	88.65	96.02	96.38
w/o L_{msc}	97.83	98.14	98.02	99.01	88.35	89.64	96.87	97.52
Ours	**98.54**	**99.08**	**98.86**	**99.87**	**89.13**	**90.21**	**97.59**	**98.23**

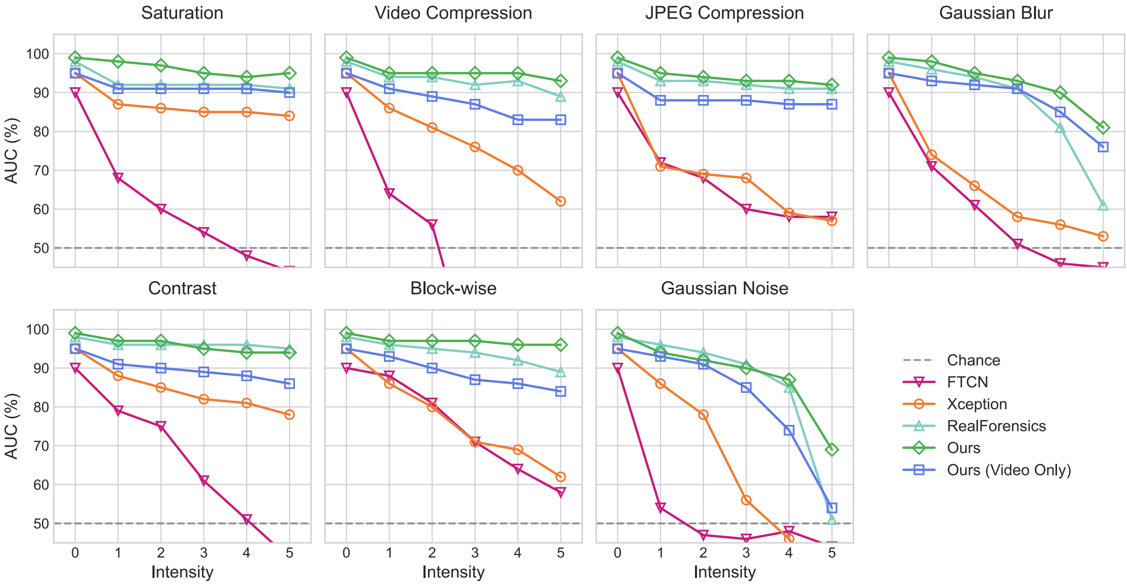

Fig. 4. Robustness evaluation against unseen visual perturbations. AUC scores (%) comparing our method with other detectors (FTCN, Xception, RealForensics) are presented as a function of perturbation intensity. Seven distinct types of perturbations commonly encountered in the wild are evaluated individually: Block-wise distortion, Saturation change, Gaussian Blur, Gaussian Noise, Video Compression, JPEG Compression, and Contrast change. Each perturbation is applied at six intensity levels (0 to 5). Higher AUC indicates superior detection performance.

AST's self-attention mechanism more effectively models long-term dependencies and semantic coherence in audio. Ablation results show that removing the AST module leads to an average decrease of 9.06% ACC and 9.22% AUC, with a particularly significant AUC drop of 9.77% on the KoDF dataset. The effectiveness of AST derives from its capability to: 1) capture subtle semantic patterns in speaker speech that often appear unnatural in synthesized speech; and 2) establish semantic consistency verification between audio features and corresponding visual content.

Impact of Loss Functions. Our method employs a combination of three loss functions: momentum-based multimodal contrastive loss (L_{mac}), cross-modal consistency loss (L_{ocs}), and modality-specific contrastive loss (L_{msc}). Among these, L_{mac} effectively distinguishes between real and forged samples' cross-modal relationships by maintaining a dynamically updated negative sample queue. Experimental results demonstrate that each of these three loss functions significantly contributes to the final performance. Specifically, removing L_{mac} results in an average decrease of 3.5% ACC and 3.74% AUC; eliminating L_{ocs} causes an average reduction of 1.79% ACC and 1.56% AUC; while excluding L_{msc} leads to a decrease of 0.84% ACC and 0.79% AUC. These findings indicate that L_{mac} contributes most substantially to model performance, followed by L_{ocs} and L_{msc}.

Overall Performance. Our complete method achieves optimal performance across all four datasets, with average ACC and AUC reaching 96.03% and 96.85%, respectively.

Particularly on the DF-TIMIT (HQ) dataset, our approach attains exceptional results of 98.86% ACC and 99.87% AUC, demonstrating the effectiveness and robustness of our proposed multimodal deepfake detection framework. Compared to unimodal methods, our comprehensive approach improves performance by an average of 19.27% ACC and 19.54% AUC. This significant enhancement can be attributed to the synergistic effect of TAViT, AST, facial UV mapping, and the three specially designed loss functions working in concert.

E. Robustness to Unseen Perturbations

Evaluating the robustness of forgery detectors against post-processing operations is crucial, as manipulated videos often undergo various quality-degrading transformations during online redistribution, complicating detection. To address this, we conducted a rigorous assessment of our proposed method's resilience. We adopted a standard evaluation framework [61], subjecting test samples to seven prevalent visual perturbations: Block-wise distortion, Saturation adjustment, Gaussian Blur, Gaussian Noise, Video Compression, JPEG Compression, and Contrast modification. The impact of these distortions was examined across six intensity levels (0-5), using the Area Under the Curve (AUC) as the primary performance metric. Our model's performance was benchmarked against three widely recognized detectors: FTCN [59], Xception [57], and the state-of-the-art RealForensics [61].

Fig. 4 reveals distinct robustness profiles for each detector. While all methods exhibit performance degradation as perturbation intensity increases, our approach demonstrates notable resilience. It consistently maintains a significant performance advantage over FTCN and Xception across the vast majority of tested scenarios. FTCN, in particular, shows rapid performance collapse under several distortions like compression and contrast changes. The comparison with RealForensics highlights the strengths of our method. Our method achieves superior robustness against Block-wise distortion, Saturation, JPEG Compression, and Contrast manipulations, maintaining higher AUC scores especially as intensity levels rise. Although RealForensics exhibits marginal advantages under specific visual corruptions like Gaussian Blur compared with our model without audio features, it operates solely on the visual stream. In contrast, our multimodal approach leverages both audio and visual information, maintaining highly competitive overall performance and broader resilience against various artifacts. This overall strong performance, despite various challenging visual corruptions, underscores the effectiveness and practical potential of our detector for real-world deployment where pristine source videos cannot be guaranteed.

F. Computational Cost Analysis

To address practical deployment considerations, we analyze the computational costs of our proposed FUME framework, focusing on model complexity and inference efficiency. All experiments are conducted on NVIDIA RTX 4080 GPU to ensure fair comparison.Table IV presents a comprehensive comparison of model parameters, storage requirements, and inference speed across different methods.

TABLE IV
COMPUTATIONAL COST COMPARISON OF DIFFERENT DEEPFAKE DETECTION METHODS.

Method	Modality	Parameters	Inference Time (ms)
Capsule [56]	$\mathcal{V}$	3.90M	33.98
Xception [57]	$\mathcal{V}$	39.53M	24.35
LipForensics [47]	$\mathcal{V}$	36.23M	28.73
EmoForen [62]	$\mathcal{AV}$	42.73M	41.67
MDS [63]	$\mathcal{AV}$	47.36M	48.32
AVA-CL [7]	$\mathcal{AV}$	42.26M	38.24
MCL [4]	$\mathcal{AV}$	48.31M	47.30
Ours	$\mathcal{AV}$	56.12M	43.86

The proposed FUME framework, with a total parameter count of 56.12M, aligns with the scale of SOTA multimodal deepfake detection methods while demonstrating robust detection performance. Its inference efficiency is characterized by a single-frame processing time of 43.86 ms, corresponding to a processing rate of 22.8 FPS. This metric reflects the computational cost of its comprehensive multimodal pipeline, which integrates synchronized audio-visual feature extraction, temporal modeling, and cross-modal fusion.

In the context of multimodal methods, FUME's processing rate (22.8 FPS) places it in the middle of the performance spectrum. Specifically, it is slightly slower than EmoForene (24.0 FPS) and MCL (21.1 FPS) but significantly faster than MDS (20.7 FPS), with its throughput falling within the typical range of multimodal architectures. Notably, while it does not outpace AVA-CL (26.2 FPS), its speed remains competitive relative to methods with similar architectural complexities, balancing the computational demands of multi-modal processing with practical feasibility.

FUME's computational design prioritizes detection accuracy over raw processing speed, making it particularly suitable for deepfake detection scenarios where precision is critical. The framework's 22.8 FPS processing rate is well-suited for offline deepfake detection tasks, social media content screening, and security surveillance applications where comprehensive analysis takes precedence over real-time processing. The integration of facial UV maps via 3DDFA, momentum contrastive learning mechanisms, and One-Class Softmax loss introduces computational complexity that directly translates to enhanced robustness against sophisticated deepfake generation techniques including face-swapping, face reenactment, and audio-visual manipulation. For deployment contexts requiring higher throughput, the modular architecture enables potential optimizations through component parallelization and hardware acceleration, while maintaining the core advantages of cross-modal semantic alignment and texture-aware spatiotemporal modeling.

V. Conclusion and Limitations

In this work, we propose FUME, a robust multimodal deepfake detection framework that effectively addresses key challenges through the innovative integration of facial UV maps and momentum contrastive learning. The framework employs a Texture-Aware Video Transformer (TAViT) for precise spatio-temporal modeling and an Audio Spectrogram Transformer (AST) for multi-scale spectral analysis, while our momentum contrastive alignment mechanism enhances cross-modal consistency and One-class softmax loss improves generalization capability against unseen deepfake generation techniques. Extensive experiments across multiple benchmark datasets validate the effectiveness of our approach, with FUME achieving 97.59% accuracy and 98.23% AUC on the challenging KoDF dataset, surpassing existing state-of-the-art methods by 1.91% and 2.99% respectively.

Despite these promising results, several limitations warrant careful consideration. The computational complexity of our framework, stemming from the integration of facial UV map generation, dual-transformer architectures, and momentum contrastive learning, results in substantial overhead with 264M parameters and 97ms inference time per sample. This computational burden may limit deployment in resource-constrained environments or real-time applications where processing speed is critical. Additionally, our framework's dependency on accurate facial localization through S^3FD introduces potential vulnerabilities when faces are partially occluded, heavily compressed, or deliberately obscured by adversarial modifications.

Most critically, the threat of adaptive attacks represents a fundamental challenge that extends beyond our current evaluation framework. Our comprehensive assessment primarily focuses on existing deepfake generation methods and may not adequately represent sophisticated adaptive attacks specifically engineered to circumvent detection mechanisms. The multimodal nature of FUME, while providing enhanced detection capabilities, simultaneously creates multiple attack surfaces that could be systematically exploited. Adversaries could potentially reverse-engineer our momentum contrastive learning mechanism to generate deepfakes that maintain consistent cross-modal representations while remaining synthetic, or develop advanced texture synthesis techniques that produce natural-looking UV maps to bypass our texture-aware detection components. Furthermore, sophisticated attackers could ensure tighter audio-visual synchronization through advanced lip-sync techniques and voice cloning methods, effectively exploiting the cross-modal consistency that our system relies upon for detection. Future research should prioritize the development of robust evaluation protocols against adaptive attacks and design dynamic defense mechanisms that can evolve with emerging attack strategies.

References

[1] Y. Nirkin *et al.*, "Fsgan: Subject agnostic face swapping and reenactment," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 7184–7193.

[2] M. A. Raza, K. M. Malik, and I. Ul Haq, "Holisticdfd: Infusing spatiotemporal transformer embeddings for deepfake detection," *Information Sciences*, vol. 645, p. 119352, Oct. 2023.

[3] S. Usmani, S. Kumar, and D. Sadhya, "Efficient deepfake detection using shallow vision transformer," *Multimedia Tools and Applications*, Jun. 2023.

[4] X. Liu, Y. Yu, X. Li, and Y. Zhao, "Mcl: Multimodal contrastive learning for deepfake detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 4, pp. 2803–2813, 2024.

[5] K. e. a. Tian, "Illumination enlightened spatial-temporal inconsistency for deepfake video detection," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2024.

[6] S. K. e. a. Datta, "Exposing lip-syncing deepfakes from mouth inconsistencies," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2024.

[7] Y. Zhang, W. Lin, and J. Xu, "Joint audio-visual attention with contrastive learning for more general deepfake detection," *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2023.

[8] C. Yu, S. Jia, X. Fu, J. Liu, J. Tian, J. Dai, X. Wang, S. Lyu, and J. Han, "Explicit correlation learning for generalizable cross-modal deepfake detection," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2024.

[9] K. Lee, Y. Zhang, and Z. Duan, "A multi-stream fusion approach with one-class learning for audio-visual deepfake detection," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*. West Lafayette, IN, USA: IEEE, 2024, pp. 1–6.

[10] P. Kwon, J. You, G. Nam, S. Park, and G. Chae, "Kodf: A large-scale korean deepfake detection dataset," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 10 724–10 733.

[11] B. Dolhansky *et al.*, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint*, 2020. [Online]. Available: https://arxiv.org/abs/2006.07397

[12] Y. Gong, Y.-A. Chung, and J. Glass, "Ast: Audio spectrogram transformer," in *Interspeech (INTERSPEECH)*, 2021, pp. 571–575.

[13] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and S. Z. Li, "Towards fast, accurate and stable 3d dense face alignment," in *European Conference on Computer Vision (ECCV)*, 2020.

[14] Y. Dahdouh, A. A. Boudhir, and M. B. Ahmed, "A new approach using deep learning and reinforcement learning in healthcare: Skin cancer classification," *International Journal of Electrical and Computer Engineering Systems*, vol. 14, no. 5, p. Art. no. 5, Jun. 2023.

[15] D. Cozzolino, A. Pianese, M. Nießner, and L. Verdoliva, "Audio-visual person-of-interest deepfake detection," in *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 943–952.

[16] T. Wang, H. Cheng, K. P. Chow, and L. Nie, "Deep convolutional pooling transformer for deepfake detection," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 19, no. 6, pp. 179:1–179:20, May 2023.

[17] Y.-J. Heo, W.-H. Yeo, and B.-G. Kim, "Deepfake detection algorithm based on improved vision transformer," *Applied Intelligence*, vol. 53, no. 7, pp. 7512–7527, Apr. 2023.

[18] B. Liang, Z. Wang, B. Huang, Q. Zou, Q. Wang, and J. Liang, "Depth map guided triplet network for deepfake face detection," *Neural Networks*, vol. 159, pp. 34–42, Feb. 2023.

[19] X. Li, R. Ni, P. Yang, Z. Fu, and Y. Zhao, "Artifacts-disentangled adversarial learning for deepfake detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 4, pp. 1658–1670, Apr. 2023.

[20] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang, and Y. Zhao, "Audio deepfake detection: A survey," *arXiv preprint arXiv:2308.14970*, 2023.

[21] X. Liu, X. Wang, M. Sahidullah, J. Patino, H. Delgado, T. H. Kinnunen, J. Yamagishi, N. Evans, K. A. Lee, and A. Nautsch, "Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2507–2522, 2022.

[22] W. Ge, M. Panariello, J. Patino, M. Todisco, and N. Evans, "Partially-connected differentiable architecture search for deepfake and spoofing detection," in *Interspeech (INTERSPEECH)*, 2021, pp. 4339–4343.

[23] C. Wang, J. Yi, J. Tao, C. Y. Zhang, S. Zhang, and X. Chen, "Detection of cross-dataset fake audio based on prosodic and pronunciation features," in *Proceedings of Interspeech (INTERSPEECH)*, 2023.

[24] X. Wang and J. Yamagishi, "Investigating self-supervised front ends for speech spoofing countermeasures," in *Odyssey 2020 The Speaker and Language Recognition Workshop (ODYSSEY)*, 2021, pp. 295–302.

[25] Z. Wu, R. K. Das, J. Yang, and H. Li, "Light convolutional neural network with feature genuinization for detection of synthetic speech attacks," in *Proceedings of Interspeech (INTERSPEECH)*, 2020, pp. 1101–1105.

[26] X. Li, N. Li, C. Weng, X. Liu, D. Su, D. Yu, and H. Meng, "Replay and synthetic speech detection with res2net architecture," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6354–6358.

[27] J.-w. Jung, H.-S. Heo, H. Tak, H.-j. Shim, J. S. Chung, B.-J. Lee, H.-J. Yu, and N. Evans, "Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6367–6371.

[28] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with rawnet2," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6369–6373.

[29] S. Ding, Y. Zhang, and Z. Duan, "Samo: Speaker attractor multi-center one-class learning for voice anti-spoofing," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[30] X. Zhang, J. Yi, J. Tao, C. Wang, and C. Y. Zhang, "Do you remember? overcoming catastrophic forgetting for fake audio detection," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

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

[32] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, and L. Verdoliva, "Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. Vancouver, BC, Canada: IEEE, Jun. 2023, pp. 20 606–20 615.

[33] Y. Wang, K. Yu, C. Chen, X. Hu, and S. Peng, "Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7278–7287.

[34] M. A. Manzoor, S. AlBarri, Z. Xian, Z. Meng, P. Nakov, and S. Liang, "Multimodality representation learning: A survey on evolution, pretraining and its applications," *arXiv preprint*, 2023.

[35] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[36] Y. Li, M. Chang, and S. Lyu, "In ictu oculi: Exposing ai-generated fake face videos by detecting eye blinking," *arXiv preprint*, 2018.

[37] Z. Qian, P. Baaquie, Y. Gan, S. Shamsi, D. Hou, J. Xu *et al.*, "Learning to detect manipulated facial images," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[38] B. Shi, W. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *International Conference on Learning Representations (ICLR)*, 2022.

[39] J. S. Chung, A. Jamaludin, and A. Zisserman, "You said that?: Synthesising talking faces from audio," *International Journal of Computer Vision*, 2019.

[40] A. Jaiswal, S. Bhatnagar, R. Patel, M. Avila-Herrera, and A. F. Martins, "Detecting ai-synthesized speech using bispectral analysis," in *Conference on Computer Vision and Pattern Recognition Workshop on Media Forensics (CVPRW)*, 2020.

[41] Y. Zhuge, D. Gao, D. Fan, L. Jin, B. Chen, H. Zhou, M. Qiu, and L. Shao, "Kaleido-bert: Vision-language pre-training on fashion domain," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[42] M. Mittal, M. Vatsa, and R. Singh, "Multimodal fusion for deepfake detection," in *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.

[43] Y. Lin, X. Han, N. Qiu, J. Deng, and G. Yang, "Av-transformer: Audiovisual inconsistency detection for fake videos," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[44] Y. Wang, M. Yasunaga, H. Ren, S. Wada, and J. Leskovec, "Vqa-gnn: Reasoning with multimodal semantic graph for visual question answering," *arXiv preprint*, 2022.

[45] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba, "The sound of pixels," in *European Conference on Computer Vision (ECCV)*, 2018.

[46] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong, "Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[47] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5039–5049.

[48] S. Kumar, S. Garg, and H. A. Moreno, "Protecting world leaders against deep fakes using facial, audio, and cross-modal consistency checks," in *International Conference on Multimedia (MM)*, 2020.

[49] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge dataset," *arXiv preprint*, 2020.

[50] D. Agarwal, D. Yadav, N. Kohli, R. Singh, and M. Vatsa, "Detecting deepfakes with metric learning," in *International Conference on Pattern Recognition (ICPR)*, 2020.

[51] L. Chen, X. Cui, C. Li, Y. Zhang, V. Venu, and J. Liu, "Detecting forged audio via spectro-temporal attention network," in *International Conference on Multimedia and Expo (ICME)*, 2021.

[52] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S3fd: Single shot scale-invariant face detector," in *International Conference on Computer Vision (ICCV)*, 2017, pp. 192–201.

[53] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," *arXiv preprint arXiv:1912.02781*, 2019.

[54] R. Shao, T. Wu, J. Wu, L. Nie, and Z. Liu, "Detecting and grounding multi-modal media manipulation and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5556–5574, 2024.

[55] Y. Zhang, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[56] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2307–2311.

[57] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 1–11.

[58] D. Wodajo and S. Atnafu, "Deepfake video detection using convolutional vision transformer," *arXiv preprint arXiv:2102.11126*, 2021.

[59] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021, pp. 15 044–15 054.

[60] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-ray for more general face forgery detection," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5001–5010.

[61] A. Haliassos, R. Mira, S. Petridis, and M. Pantic, "Leveraging Real Talking Faces via Self-Supervision for Robust Forgery Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14 950–14 962.

[62] T. Mittal *et al.*, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in *ACM Multimedia (MM)*. ACM, 2020, pp. 2823–2832.

[63] K. Chugh *et al.*, "Not made for each other-audio-visual dissonance-based deepfake detection and localization," in *ACM Multimedia (MM)*, 2020, pp. 439–447.

[64] R. Korshunov and S. Marcel, "Deepfaketimit: A deepfake video dataset for detection," arXiv preprint arXiv:1812.08685, 2020, accessed: [Insert Access Date].

[65] W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, X. Cao, and K. Ren, "Avoid-df: Audio-visual joint learning for detecting deepfake," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 2015–2029, 2023.

[66] X. Liu, Y. Yu, X. Li, and Y. Zhao, "Mcl: Multimodal contrastive learning for deepfake detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 4, pp. 2803–2813, 2024.