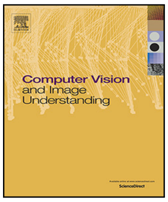
Computer Vision and Image Understanding 248 (2024) 104133

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu



# Audio–visual deepfake detection using articulatory representation learning

Yujia Wang, Hua Huang [*]

*Beijing Normal University, Beijing, 100875, PR China*

ARTICLE INFO

*Keywords:*
Deepfake detection
Audio–visual
Articulatory representation

ABSTRACT

Advancements in generative artificial intelligence have made it easier to manipulate auditory and visual elements, highlighting the critical need for robust audio–visual deepfake detection methods. In this paper, we propose an articulatory representation-based audio–visual deepfake detection approach, *ART-AVDF*. First, we devise an audio encoder to extract articulatory features that capture the physical significance of articulation movement, integrating with a lip encoder to explore audio–visual articulatory correspondences in a self-supervised learning manner. Then, we design a multimodal joint fusion module to further explore inherent audio–visual consistency using the articulatory embeddings. Extensive experiments on the DFDC, FakeAVCeleb, and DefakeAVMiT datasets demonstrate that *ART-AVDF* obtains a significant performance improvement compared to many deepfake detection models.

## 1. Introduction

Deepfake aims to produce highly realistic fake videos where a person is manipulated by artificial intelligence technology. Editing tools (*e.g.*, Sora[1]) and Text-to-Speech (TTS) techniques, have seen significant advancements, leading to the widespread creation and manipulation of synthetic content. To mitigate the impact of such multimedia in non-entertainment and critical contexts, such as surveillance, emergency services, and scientific research, numerous studies are focusing on detecting manipulated content.

For unimodal deepfake detection approaches, the core strategy involves, for example, identifying inconsistencies in facial features (Hu et al., 2022), lip movements (Haliassos et al., 2021), low-level texture features (Zhao et al., 2021) in manipulated videos, or distinguishing genuine utterances from fake audio clips (Yi et al., 2022, 2023). These approaches face challenges in addressing real-world scenarios that involve manipulations across multiple modalities (Zou et al., 2024) and in collecting datasets that fully capture all possible manipulations in the wild (Feng et al., 2023). Moreover, researchers have induced audio-prompt visual content generation (Oh et al., 2019; Bigioi et al., 2024), and corruptions can occur in both visual and auditory channels. This has led to the development of audio–visual deepfake detection tasks.

The key insight of audio–visual deepfake detection is that the information of the two channels can synergistically collaborate and enhance each other, similar to how humans precept multimodal information (Hershey and Movellan, 1999). There are two ways for the joint learning of multimodal information: (i) ensemble learning, *i.e.* analyzing multimodal data separately and making decisions based on similarity scores, *e.g.*, VFD (Cheng et al., 2023) and POI-Forensics (Cozzolino et al., 2023); (ii) applying different learning strategies to map unimodal features, *e.g.*, Avoid-DF (Yang et al., 2023). However, these strategies sometimes prove unsatisfactory for various scenarios. First, the ensemble learning strategy may fail when both visual and auditory data are manipulated (Cheng et al., 2023). Second, they overlook the potential insights speech data could offer beyond latent representations and auxiliary lip motion data. The rich articulation in human speech significantly enhances the understanding and analysis of audio–visual correspondence. Given this background, we delve into more comprehensive motion patterns associated with speech and desire to explore a potential solution for representation learning of articulation movement.

In this paper, we propose to compile physiological-level audio–visual correlations, i.e., articulatory-related representations, into popular audio–visual deepfake detection mechanisms. Our key insight is that articulatory information implicitly involved in the video can serve as strong learning signals. For instance, speeches are produced by the coordinated movements of various articulatory organs. As shown in Fig. 1, with variations in organs' shapes, sizes, and movement patterns, different speakers possess their own distinctive voices and speaking styles (Fant, 1960). We also 'bind' auditory and visual information; *i.e.* lip motions have been linked to articulatory movement in real videos.

Building on this insight, we introduce *ART-AVDF*, an audio–visual deepfake detection model that seeks to enhance performance by utilizing articulatory representations. As shown in Fig. 1, *ART-AVDF* consists of two modules, *i.e.* articulatory representation learning module (*ART*) and audio–visual deepfake detection module (*AVDF*). The *ART* module

[*] Corresponding author.
*E-mail address:* huahuang@bnu.edu.cn (H. Huang).

[1] https://openai.com/sora

https://doi.org/10.1016/j.cviu.2024.104133
Received 31 May 2024; Received in revised form 19 August 2024; Accepted 20 August 2024
Available online 22 August 2024

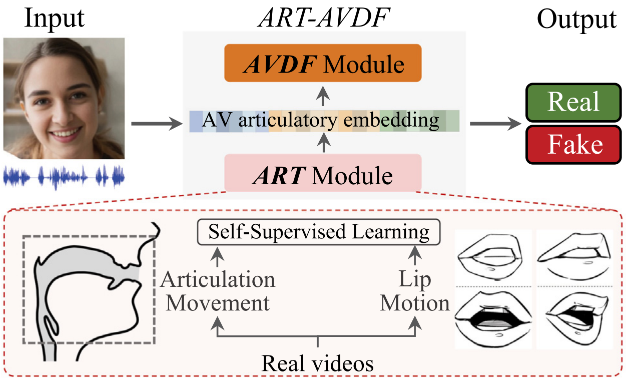


**Fig. 1.** Overview of our proposed *ART-AVDF*, which consists of two modules, *i.e. ART* module, aiming to analyze audio–visual articulatory correspondences from real videos; *AVDF* module, performing audio–visual deepfake detection with articulatory embedding.

aims to capture the implicit audio–visual articulatory correspondences, *i.e.* the physiological significance of articulation movement and lip motion. Specifically, we propose two-stream encoders to encode auditory and visual information respectively, and perform self-supervised training on real videos. The *AVDF* module aims to analyze audio–visual consistencies with the help of articulatory representations. A cross-modal attention model is designed to fuse multi-modal features. We leverage the cross-entropy loss and contrast learning strategy for forgery detection.

To evaluate the effectiveness of our approach, we conducted experiments on videos containing manipulated speeches and human faces, *i.e.* DFDC (Dolhansky et al., 2020), FakeAVCeleb (Khalid et al., 2021b), and DefakeAVMiT (Yang et al., 2023). Experimental results demonstrate that our model achieves improved performance. In general, our contributions could be summed up as follows:

- We propose a novel framework, *ART-AVDF*, that utilizes articulatory representation for accurate audio–visual deepfake detection.
- In *ART* module, an auditory encoder and a lip encoder are designed to perform audio–visual articulatory representation learning using the self-supervised learning strategy. In *AVDF* module, we utilize the frozen encoders in *ART* module to obtain articulatory embeddings and fuse them with unimodal features, leading to better audio–visual analysis for deepfake detection.
- We conduct extensive experiments on different datasets comprising visual, audio, and audio–visual deepfakes to signify the effectiveness of our *ART-AVDF*.

The rest of the paper is organized as follows. Related works are briefly reviewed in Section 2. The proposed framework for using articulatory representation for audio–visual deepfake detection is presented in Section 3. Section 4 shows the experimental results and corresponding discussions. The conclusion is drawn in Section 5.

## 2. Related work

In this section, we first give a concise overview of audio–visual deepfake detection. Then, we give a brief discussion of audio–visual consistency and articulatory representation learning that pertains closely to our work.

### 2.1. Audio–visual deepfake detection

Early works primarily focused on either audio or visual cues independently — detecting auditory manipulations by finding inconsistencies in reverberation (Malik and Farid, 2010); detecting visual forgery through biological artifacts, *e.g.*, unnatural eye blinking (Liy and InIctuOculi, 2018) and inconsistent head pose (Yang et al., 2019). Please refer to Nguyen et al. (2022) for a detailed survey. Recently, studies have demonstrated the effectiveness of multimodal deep learning-based approaches that integrate both visual and auditory data. This advancement is significant in response to the growing sophistication of deepfake technologies that manipulate both audio and visual elements to create realistic fake content.

The current efforts can be divided into two categories. First, ensemble learning — researchers leveraged the ensemble of audio and visual networks, *i.e.* fusing features or scores from two modalities to make the final real/fake decision (Ilyas et al., 2023). For example, Chugh et al. (2020) searched for inconsistencies between audio and visual information by training a modality dissonance score. Zhou and Lim (2021) proposed to learn and exploit the intrinsic synchronization between video and audio. Building on this strategy, Cheng et al. (2023) analyzed the intrinsic correlation of facial and audio and enhanced the traditional contrastive loss to align the objective between generic and deepfake datasets. Similarly, Cozzolino et al. (2023) followed the contrastive learning paradigm to learn the moving-face and audio segment embedding that are most discriminative for each identity. However, this kind of framework has been proven to be non-optimal sometimes when the videos contain carefully manipulated soundtrack and corresponding visuals.

On the other hand, many efforts have been made to explore joint audio–visual learning (Raza and Malik, 2023; Muppalla et al., 2023; Zhang et al., 2024), as the two modalities are often intertwined. Yang et al. (2023) introduced AVoiD-DF, an encoder–decoder network designed to fuse audio–visual features and jointly learn the corresponding inherent relationships, along with a cross-modal classifier to detect manipulation with inter-modal and intra-modal disharmony. Cai et al. (2023) leveraged a 3D convolutional neural network-based architecture to capture multimodal manipulations and provided precise boundaries of fake segments in videos. Moreover, the self-supervised learning technique has been wildly applied in audio–visual deepfake detection task (Korbar et al., 2018; Afouras et al., 2020; Zong et al., 2023). For example, Yu et al. (2023) and Feng et al. (2023) utilized self-supervised learning to eliminate audio–visual gaps and provide inherent correspondences. Zhao et al. (2022) desired to learn mouth motion representations by promoting similarity in paired video and audio representations.

Due to the gap between auditory and visual modalities, uncovering potential correlations and subtle inconsistencies between the two modalities is critical and needs to be further explored. Inspired by these works, we desired to further explore whether the significant physiological information involved during someone's talking could help improve the deepfake detection performance.

### 2.2. Articulatory representation learning

When humans speak, the coordination of various organs involved is a complex physiological process, *i.e.* the airflow from the lung vibrates through the vocal cord and then is shaped by the resonant cavity to produce sounds. With these organs' variations of shapes, sizes, and movement patterns, different speakers have their unique voices and prosody (Fant, 1960).

As early as the 1790s, Kempelen et al. (1791) developed a speaking machine, simulating human speech production by using a bellow to replicate lung function, a flute to mimic the vocal cords, and a tube to stand in for the mouth. After the 1970s, the articulatory system's physiological structure was further described through 2D geometry (Perkell, 1974; Lindblom and Sundberg, 1971) and 3D structural (Wilhelms-Tricarico, 1995; Gérard et al., 2006; Fang and Dang, 2006). For example, the transfer function method (Fant, 1960) and the impedance phase shift method (Lin and Fant, 1989) have been proposed to simulate the relationship between the vocal tract area function and the vowel formant frequency. These models enabled the simulation of the displacement and deformation of soft tissues (Birkholz, 2013).

With the most advanced medical equipment (*e.g.*, magnetic resonance imaging (MRI), ultrasound scan, and electromagnetic articulography (EMA)), researchers could capture real-time data on the movements of articulatory organs, leading to practical applications, like Text-to-Speech synthesis (TTS) (Narayanan et al., 2014; Eshky et al., 2019; Tiede et al., 2017; Richmond et al., 2011). And GMM (Toda et al., 2004), HMM (Ling et al., 2009), and deep learning-based techniques (Zhang et al., 2021; Wu et al., 2023b; Yu et al., 2021) have been employed for articulatory representation learning.

In this paper, we desire to explore the intuitive power of the articulation movement. Also, lip movement is a vital aspect of human communication, enhancing accuracy in sharing messages (Scherer, 1992). Therefore, we propose an audio–visual articulatory representation learning module to enhance the performance of multi-modal modeling.

## 3. Approach

In this section, we will begin by describing the problem formulation and providing an overview of our proposed *ART-AVDF* (Section 3.1). We detail the audio–visual articulatory representation learning module (*ART*) in Section 3.2. We also elaborate on the audio–visual deepfake module (*AVDF*) using the learned representations in Section 3.3, including the explanation of the total losses used for training and how the model is used to make the final real/fake decisions.

### 3.1. Problem formulation and overview

The goal of *ART-AVDF* is to detect if the input audio–visual media is a deepfake video. We denote the input videos containing human talking as $\mathbb{D} = \{a_i, v_i, \mathbf{y}\}_i^N, \mathbf{y} = \{y_i, y_i^v, y_i^a\} \in \{0, 1\}$, where $a_i$ and $v_i$ are the respective auditory and visual data and are sequences of sampled waveform digits and video frames. $\mathbf{y}$ are labels for the inputs. $N$ is the number of audio–visual pairs and $i$ represents the sample index for a clearer explanation. The *AVDF* network that makes the prediction is denoted as $F(x)$, which includes the audio–visual articulatory features $f_{av}$ extracted from the pre-trained *ART*; the unimodal features extracted from the input video ($f_v$) and audio ($f_a$) are processed through an audio–visual fusion module conditioned on $f_{av}$; the classification layer $F_\phi$ maps feature representations to labels. Overall, we propose to utilize auditory and visual data in cooperation with articulatory representations, to further capture inherent audio–visual correspondences in real video.

### 3.2. Articulatory representation learning

To capture the nuances of audio–visual articulation shown in the video, we first extract auditory and visual articulatory features, respectively. As shown in Fig. 2, for the auditory stream, we propose using energy, fundamental frequency, and vocal tract variables (TVs) as intermediate features to perform audio feature extraction. These articulatory-related features have been used in speaker-dependent representation learning (Seneviratne et al., 2019). On the other hand, we crop the face frames to isolate and obtain detailed videos of the lips and design a lip encoder to extract the visual features. Finally, we conduct cross-modal articulatory similarity learning by leveraging the self-supervised learning strategy.

#### 3.2.1. Auditory stream

The articulatory representation learning module (*ART*) models human speech production using three articulatory features.

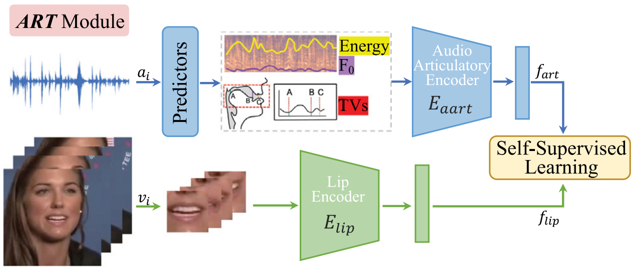


**Fig. 2.** The architecture of the *ART* module, which employs a two-stream network to explore audio–visual articulatory correspondences.

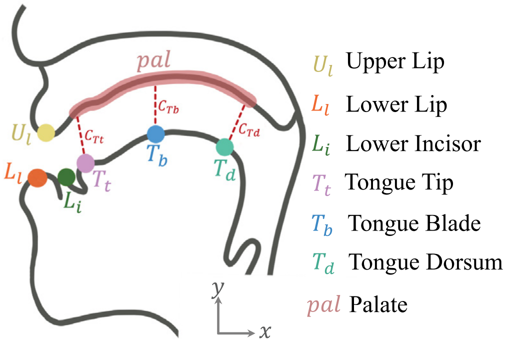


**Fig. 3.** EMA data (points) and pre-recorded plane coordinates of the palate. The red dashed lines present constriction degrees between the 3 tongue positions and the palate curve, forming vocal tract variables (TVs) along with distances between every two points.

**Energy.** At the start of speaking, the lungs expel air, powering pronunciation. The volume and velocity of expelled air are typically proportional to the energy of the sound (Motie-Shirazi et al., 2021). This process may lead the mouth to open and give rise to distinct vocal sounds. The norm of the mel spectrogram of the input audio is used to approximate the energy of the speech.

**Fundamental frequency.** The airflow impacts the vocal cords, making them vibrate. The frequency of these vibrations directly determines the fundamental frequency ($\mathrm{F}_0$)[2] of the speech (Fujisaki, 1983). We employ the JDC network (Kum and Nam, 2019) to estimate the fundamental frequency ($\mathrm{F}_0$) sequence from the mel spectrogram.

**Vocal tract variables (TVs).** Our vocal tract shapes, shaped by articulatory organs from the larynx to the lips, modulate formant frequencies (Diehl, 2008), facilitating the articulation of different phonemes and voice tones (Fant, 1960). Specifically, Vocal tract shape affects formant frequencies through dynamic changes in tongue position, lip shape, jaw opening, and vocal tract length, for example (Fant, 1960; Lee et al., 2016). Following this theory, we design a TV predictor to predict real-time relative distances among vocal organs from speeches. We first provide a detailed definition of these distances, which could be calculated from the data recorded by electromagnetic articulography (EMA) technique (Gaines et al., 2021).

As shown in Fig. 3, we leverage the real-time recorded six articulatory points, *e.g.*, the tongue tip ($T_t$), tongue blade ($T_b$), etc., to define 10 distances $D$ and corresponding constriction degrees $C$. For example, the tongue tip movement can be represented by the distance between the tongue tip ($T_t$) and the lower incisor ($L_i$), which could be defined as

[2] A similar concept is 'Pitch', which describes how our ears and brains interpret the signal, yet $\mathrm{F}_0$ describes the actual physical phenomenon.

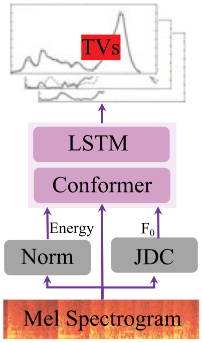


**Fig. 4.** The architecture of our proposed TV predictor.

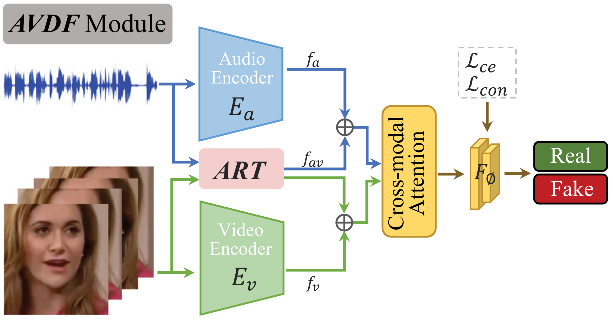


**Fig. 5.** The architecture of the *AVDF* module. The frozen *ART* module is used to provide audio–visual articulatory embedding.

$D_{\mathrm{TiTt}}[t] = \|L_i[t] - T_t[t]\|_2$, where $t$ is the timestamp. Similar definitions apply to $D_{\mathrm{TtTb}}$, indicating tongue movement; $D_{\mathrm{UlLl}}$, $D_{\mathrm{UlLi}}$, and $D_{\mathrm{LlLi}}$, for lip convexity and opening; $D_{\mathrm{LiTb}}$ and $D_{\mathrm{TbTd}}$, for tongue position and vocal tract length, respectively.

Constriction degrees, *i.e.* $C_{\mathrm{Tt}}$, $C_{\mathrm{Tb}}$, and $C_{\mathrm{Td}}$, represent the shortest distance between the 3 tongue positions and the palate curve, respectively. For example, $C_{\mathrm{Tt}}$ could be defined as $C_{T_t}[t] = \min_{x,y} \|T_t[t] - pal(x,y)\|_2$. $pal(\cdot)$ represents the point on the palate curve, with $x$ and $y$ denoting the plane coordinates within the system formed by the sagittal plane of the vocal organs. The palate location is pre-recorded as it is relatively static to the head pose while speaking (Tiede et al., 2017).

As shown in Fig. 4, we propose a TV predictor to estimate the 10-d vocal track variables (TVs) from the mel spectrogram along with the extracted energy and $F_0$. The predictor consists of five conformer blocks and one bi-LSTM layer, allowing for the integration of both global and local information and capturing short-term correlation of vocal track changes. We pre-trained our TV predictor with L1 loss on HPRC dataset (Tiede et al., 2017), which includes 7.9 h of 44.1 kHz speeches and 100 Hz EMA data recorded by 8 participants. We held out two speakers as the test set and trained our model on the other 6 speakers, achieving 0.115 on MAE and 0.987 on Corr better than that of the AAI (Wu et al., 2023a) (0.179 and 0.784). The parameters of the TV predictor will remain fixed during the subsequent model training.

**Audio articulatory encoder**. We then concatenate the predicted energy, $F_0$, and TVs and feed them into an audio articulatory encoder to extract phoneme-level features $f_{art}$. The encoder consists of a two-layer feed-forward attention layer, a three-layer AdaIN block (Huang and Belongie, 2017), and a Bi-LSTM layer followed by a linear layer.

### 3.2.2. Visual stream

The input face frames are cropped to obtain the lip video, which is then fed into the lip encoder. The encoder has stacks of 3D CNNs, which have been shown to be effective in multiple tasks involving spatial–temporal video data (Zhang et al., 2021). The lip frame is represented as $I_{lip} \in \mathbb{R}^{(C\times T\times H\times W)}$, where $C$ and $T$ denote the channel number and temporal depth, $H$ and $W$ are the height and width. The output of the lip encoder can be represented by $f_{lip} \in \mathbb{R}^{(T\times D)}$, where $T$ denotes temporal depth and $D$ denotes features dimension.

### 3.2.3. Self-supervised audio–visual articulatory similarity learning

With $f_{art}$ and $f_{lip}$ representing articulatory features from two heterogeneous modalities, we then conduct cross-modal articulatory similarity learning following the self-supervised learning strategy. We first transform $f_{lip}$ to be comparable with the auditory feature via a non-linear transformation as $f'_{lip} = g(f_{lip})$ to map corresponding representations into a common d-dimensional space before computing the contrastive loss. Specifically, $g(\cdot)$ consists of a two-layer FC, separated by Batch Normalization and ReLU, followed by global average pooling. We apply the contrastive loss to maximize the symmetric joint probability between $f_{art}$ and $f'_{lip}$, *i.e.* minimizing the negative cosine similarity:

$$\mathcal{L} = -\sum_{i=1}^{B} \log\left(\frac{e^{(v^i)^\tau (a^i)}}{e^{(v^i)^\tau (a^i)} + \sum_{(v',a')\in\mathcal{N}_i} e^{(v')^\tau (a')}}\right). \tag{1}$$

$a^i$ and $v^i$ are abbreviations for $f^i_{art}$ and $f'^i_{lip}$ with the $i$th sample, respectively. $B$ represents the size of a minibatch. $(v^i, a^i)$ serve as the positive pair and $\mathcal{N}_i = \{(v^i, a^j), (v^j, a^i) | j \in [1, \ldots, B], i \neq j\}$ constitutes the negative pairs. Similar to Alayrac et al. (2020) and Morgado et al. (2021), we employ a temperature hyper-parameter $\tau$ to control the smoothness for the distribution of pairwise similarities.

During self-supervised similarity learning, the inherent audio–visual articulatory correspondences in real videos are effectively mined.

## 3.3. Audio–visual deepfake detection

As shown in Fig. 5, in the *AVDF* module, we propose to use the pre-trained *ART* module to provide articulatory features for the audio–visual detection network. The *AVDF* network is frozen during the training of the *AVDF*, learning to better analyze audio–visual inconsistencies in deepfake videos.

### 3.3.1. Unimodal feature extraction

For the input video, We encode the visual and auditory representations from two branches. For the visual branch, we sampled the input videos at intervals of 4 ms to obtain frames $x^v \in \mathbb{R}^{3\times T_v\times H\times W}$, where $T_v$ represents the number of frames. We extract visual features $f_v \in \mathbb{R}^{d_v\times H\times W}$ from the Swin Transformer (Liu et al., 2022) and a 1D-global average pooling layer. Based on the core idea of hierarchical feature extraction, the Swin Transformer could learn effective global and local contextual representations.

On the other hand, as a spectrogram (*i.e.* 1-channel 2D images) autoencoder with reconstruction objective as self-supervision has demonstrated the effectiveness of audio–visual analysis, we transfer the audio mono-waveforms (with a sampling rate of 16 kHz) into a sequence of mel-spectrogram sample $x^a \in [0,1]^{C_a\times T_a}$, where $C_a$ denotes the mel channels and $T_a$ is the number of frames. Then, we leverage the VGGish (Hershey et al., 2017), a model designed for capturing both temporal and spectral information, to extract auditory features $f_a \in \mathbb{R}^{T_a\times d_a}$. $d_a$ is 128 as the default.

### 3.3.2. Audio–visual fusion

$F_a$ and $F_v$ extracted from different branches do not align well. So we propose to use an audio–visual cross-modal attention module to align the speeches and facial expressions by treating them as a joint attention space. We concatenate $f_v$ and $f'_{lip}$ to for $F_v$ which serves as the visual stream input for the attention module. Likewise, we concatenate $f_a$ and $f_{art}$ to form $F_a$. In the joint attention setting, attention operates simultaneously over time and space. Moreover, we utilize the multi-head attention layer to capture different aspects of the input, allowing for a more comprehensive and nuanced representation of the relationships between $F_v$ and $F_a$. At the $l$th layer, when considering a visual query, the directional attention operations can be described as:

$$\begin{aligned} v_1^{(l)} &= \mathrm{MHA}(F_a^{(l-1)}, F_v^{(l-1)}), \\ v_2^{(l)} &= \mathrm{LN}(v_1^{(l)} + F_v^{(l-1)}), \\ F_v^{(l)} &= \mathrm{LN}(f(\mathrm{Dropout}(v_2^{(l)})) + F_v^{(l-1)}). \end{aligned}$$

For the auditory direction, we modulate the auditory features $a^{(l)}$ using the visual features $v^{(l)}$ by swapping $F_v$ and $F_a$ in the above equation.

Finally, we concatenate the audio–visual representation $F_v^{(l)}$ and $F_a^{(l)}$ to form the final classification feature, *i.e.*

$$F_{va} = concat\left(F_v^{(l)}, F_a^{(l)}\right).$$

$F_{va}$ is further exploited for computing cross-modal loss functions for deepfake detection.

### 3.3.3. Learning objectives

We apply the following cross-entropy loss and contrastive loss to make the final real/fake decision.

**Cross-Entropy Loss.** The cross-entropy loss is set to obtain highly discriminative features to separate real and fake videos, based on multimodal feature $F_{va}$. Formally, the cross-entropy loss can be denoted as:

$$\mathcal{L}_{ce} = -\sum_{k=1}^{K}\left(y_k \cdot \log \frac{\exp\left(f_{va}^k\right)}{\sum_{k=1}^{K}\exp\left(f_{va}^c\right)}\right),$$

where $K = 2$ for the binary decision and $y_k$ is the one-hot vector of the ground-truth of the video, and $P_k$ is the predicted probability vector.

**Contrastive Loss.** The contrastive loss used in the deepfake detection task is designed to maximize the similarity of $F_{va}$ corresponding to fake and real labels; and minimize the similarity of $F_{va}$ among samples sharing the same label. Formally, it can be expressed as:

$$\mathcal{L}_{con} = \frac{1}{B^2}\sum_{k=1}^{B}\left[\sum_{y_k=y_g}\left(1-\mathrm{sim}\left(F_{va}^k, F_{va}^g\right)\right) + \sum_{y_k\neq y_g}\max\left(\left(\mathrm{sim}\left(F_{va}^k, F_{va}^g\right)-\alpha\right), 0\right)\right]$$

where $B$ is the batch size. Sim$(\cdot)$ is the cosine similarity.

**Total Loss.** The final loss function is composed of the two losses mentioned above, *i.e.* $\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda\mathcal{L}_{\theta_{con}}$. $\lambda$ is an optional weighting parameter for balancing the loss function.

## 4. Experiments

In this section, we first introduce the experiment setup (Section 4.1) for the audio–visual deepfake detection task, including datasets, implementation details, evaluation metrics, and compared methods. Then, we provide detailed experimental results and the corresponding analysis, including quantitative evaluation (), generalization test (Section 4.3), and ablation study (Section 4.4). We also provide mass results of the qualitative evaluation (Section 4.5).

### 4.1. Setup

**Dataset.** We leveraged three multimodal datasets commonly used for beefcake detection:

(i) DFDC (Dolhansky et al., 2020), a publicly available face swap video dataset, containing 23,654 real videos recorded from 960 identities and 104,500 fake videos. Many face swap models are used on the cropped, aligned, and resized face frames, *e.g.*, MM/NN Face Swap, NTH, FSGAN, and StyleGAN. Additionally, voice conversion techniques (*i.e.* TTS Skins) are applied to perform audio swapping on some video clips. Similar to Yu et al. (2023), we filtered the dataset to have videos without noise background sounds and finally obtained 7789 real videos and 38,055 fake videos.

(ii) FakeAVCele (Khalid et al., 2021b), containing 500 real videos and 19,500 fake cones, with an equal distribution by gender, race, and age. The fake videos were first generated by face-swapping methods (*i.e.* Faceswap and FSGAN) and voice generation methods (*i.e.* SV2TTS), and then reenacted using the Wav2Lip model. There are four types of videos in this dataset, *i.e.* RealVideo–RealAudio(RR), RealVideo–FakeAudio(RF), FakeVideo–RealAudio(FR), and FakeVideo–FakeAudio(FF). Each video lasts at least 7 s.

(iii) DefakeAVMiT (Yang et al., 2023), consists of 540 real videos and 6480 deepfake videos. 8 different visual and auditory deepfake techniques were employed, *e.g.*, Faseswap and DeepFaceLab models were used for face swapping, SV2TTS for real-time voice cloning, and Wave2Lip, EVP, and PC-AVS for audio–visual synchronization. The average duration of each clip is cropped as 4.25 s or approximately 100 visual frames.

**Compared Methods.** We conduct experiments to evaluate the following deepfake detection methods, which can be categorized into two groups:

(i) Single-Modality Models, including methods on vision and audio modality, respectively. The former includes VGG16 (Simonyan and Zisserman, 2014), Meso-4 (Afchar et al., 2018), HeadPose (Yang et al., 2019), Xception (Rossler et al., 2019), Capsule (Nguyen et al., 2019), F$^3$-Net (Qian et al., 2020). We also compared our *ART-AVDF* to models focusing on lip motion for face forgery detection (Haliassos et al., 2021). On the other hand, the methods for audio deepfake detection mainly aim to protect automatic speaker verification systems from manipulation. Similar to Yang et al. (2023) and Yu et al. (2023), we compare the effectiveness of our *ART-AVDF* in comparison to LFCC-LCNN (Monteiro et al., 2019), RawNet (Jung et al., 2019), ECAPA-TDNN (Desplanques et al., 2020), and AASIST (Jung et al., 2022). Note that, RawNet uses raw waveforms as input.

(ii) Multi-modal Model, performing deepfake detection based on both visual and auditory information. Our comparison includes VGG16-MM (Khalid et al., 2021a), MDS (Chugh et al., 2020), EmotionForensics (Mittal et al., 2020), BA-TFD (Cai et al., 2022), VFD (Cheng et al., 2023), AVoiD-DF (Yang et al., 2023), PVASS-MDD (Yu et al., 2023), and AVFakeNet (Ilyas et al., 2023). Note that, MDS and EmotionForensics ignored the fact that the auditory information could also be manipulated.

**Evaluation Metrics.** We quantify the performance by adopting standard detection metrics outlined in Zhang et al. (2024), including area under the curve (AUC) and accuracy (ACC). Particularly, unimodal methods and multi-modal methods have different detection labels (shown in Table 1).

**Implementation Details.** We resize all video frames to a size of 224 × 224. On the contrary, we extract the log Mel-Spectrogram using 64 mel filter banks over 4 ms of audio data sampled at 16 kHz. The weight for balancing the two losses is empirically set to 1. We employ SGD optimizer to optimize the model parameters with an initial learning rate of $10^{-4}$ with cosine decay. The hyperparameter temperature in InfoNCE is set to 0.1. The network is trained for 2000 epochs on the NVIDIA A800 with a batch size of 128. Following the settings outlined in Yang et al. (2023), we split the datasets with 70% for training and the rest 30% for testing.[3] Each video clip contains around 100 pairs of auditory and visual data, and each pair has two kinds of labels, *i.e.* real or fake.

[3] Models mentioned in Table 2 are trained under the same setting.

**Table 1**
Labels for different modalities. ✗ for real and ✓ for fake. "$R_V R_A$" represents the videos contain both real visual and auditory data, while "$F_V F_A$" indicates the opposite. "$R_V F_A$" and "$F_V R_A$" represent fake videos where only either the audio or visual data has been manipulated, *i.e.* fake auditory and fake visual data, respectively.

| Modality | RVRA | RVFA | FVRA | FVFA |
|---|---|---|---|---|
| Visual | ✓ | ✓ | ✗ | ✗ |
| Auditory | ✓ | ✗ | ✓ | ✗ |
| Audio–visual | ✓ | ✗ | ✗ | ✗ |

### 4.2. Quantitative evaluation

In this section, we carry out a series of experiments to evaluate the performance of our proposed *ART-AVDF*, including the comparison with unimodal deepfake detection methods (Sections 4.2.1 and 4.2.2) and multi-modal deepfake detection methods (Section 4.2.3).

#### 4.2.1. Comparison for visual deepfake detection

As shown in Table 2, the first group demonstrates the comparison results for the visual deepfake detection task. Overall, our *ART-AVDF* demonstrates considerable advantages compared to deep learning methods, *i.e.* VGG16, Meso-4, HeadPose, Xception, and Capsule, across all datasets, achieving an average improvement of over 9%. These models' results reveal significant differences among the three datasets, underscoring their limited generalization capability. $F^3$-Net achieves relatively better detection results among these models, indicating that the frequency domain effectively detects deepfakes. LipForensics, which focuses on high-level semantic irregularities in lip movements, faces challenges when lip movement is simulated by high-precision lip-sync techniques. To address this, our *ART-AVDF* takes the articulation movement into account to improve the lip motion modeling, leading to an improvement of 6.2% on average.

We adapt our framework for visual deepfake detection by removing the audio branch in the *AVDF* module and utilizing the pre-trained *ART* module to provide further visual information from the visual encoder. Overall, *ART-VDF* outperforms most models (except for $F^3$-Net) across all three datasets, demonstrating the effectiveness of our *ART* module, which could also serve as a plug-in component for other models. Also, the performance of *ART-VDF* is lower than that of *ART-AVDF*, which underscores the potential advantages of jointly analyzing both audio and visual modality.

#### 4.2.2. Comparison for audio deepfake detection

As shown in the second group of Table 2, many attempts have been proposed to resist deepfake audio attacks and protect automatic speaker verification systems. Overall, all models achieve average ACC and AUC below 75% on the DefakeAVMit dataset, and below 70% on the other two datasets. Similar to the experiment for visual deepfake detection, we also adapt our framework for audio deepfake detection by removing the visual branch, *i.e. ART-ADF*, resulting in an average of 10% enhancement across all datasets. This significant improvement could be attributed to the fact that articulatory representations would not be influenced by the tremendous noise in real-world videos, thereby reducing the impact on speaker verification performance. Also, detection models using single-modal information perform worse than those using multi-modal data; the absence of essential visual (or auditory) information may lead to wrong decisions, much like how humans perceive the world, where every sensory input holds importance.

#### 4.2.3. Comparison for audio–visual deepfake detection

Recently, researchers have paid more attention to multi-modal approaches for deepfake detection, *i.e.* analyzing both auditory and visual stimuli. For example, EmotionForensics analyzes the emotion similarity between the two modalities. However, due to the inconsistent emotion expressions in real videos, EmotionForensics achieves an average score of 82.5% (ACC) and 85.4% (AUC) (shown in the third group in Table 2). When considering the manipulated auditory information rather than exploiting it as an additional supervision signal, EmotionForensics and MDS may fail. To tackle this challenge, recent attempts, *e.g.*, BA-TFD, AVFakeNet, VFD, AVoiD-DF, and PVASS-MDD have been developed to further analysis the audio–visual correspondences and enhance the models' ability to detect and analyze fake or manipulated data with greater accuracy. Our proposed *ART-AVDF* outperforms EmotionForensics, MDS, BA-TFD, VFD, and AVFakeNet at an average of 10.8% (ACC) and 9.8% (AUC) on all three datasets and achieves slightly higher scores than that of AVoid-DF, *i.e.* only a 5.8% higher for ACC and 4.1% for AUC. PVASS-MDD achieves the highest scores during the experiments, *i.e.* over 90% across all datasets for both ACC and AUC. Our results are comparable to those of PVASS-MDD. Specifically, *ART-AVDF* achieves

**Table 2**
Performance (%) of our *ART-AVDF* and other deepfake detection models on DFDC, FakeAVCele, and DefakeAVMiT dataset.

| Model | Modality | | DFDC | | FakeAVCele | | DefakeAVMiT | |
|---|---|---|---|---|---|---|---|---|
| | Visual | Audio | ACC | AUC | ACC | AUC | ACC | AUC |
| VGG16 (Simonyan and Zisserman, 2014) | ✓ | – | 53.2 | 59.4 | 52.9 | 54.1 | 46.0 | 49.2 |
| Meso-4 (Afchar et al., 2018) | ✓ | – | 71.7 | 75.3 | 57.3 | 60.9 | 80.1 | 82.5 |
| HeadPose (Yang et al., 2019) | ✓ | – | 51.4 | 55.9 | 45.6 | 49.2 | 50.6 | 52.2 |
| Xception (Rossler et al., 2019) | ✓ | – | 46.5 | 49.9 | 67.9 | 70.5 | 49.4 | 51.0 |
| Capsule (Nguyen et al., 2019) | ✓ | – | 50.2 | 53.3 | 68.8 | 70.9 | 68.5 | 71.4 |
| $F^3$-Net (Qian et al., 2020) | ✓ | – | 73.2 | 75.4 | 83.6 | 84.5 | 79.5 | 81.7 |
| LipForensics (Haliassos et al., 2021) | ✓ | – | 71.3 | 73.5 | 80.1 | 82.4 | 73.1 | 77.2 |
| ***ART-VDF*** | ✓ | – | 81.2 | 85.4 | 80.7 | 84.3 | 80.6 | 82.7 |
| VGG16 (Simonyan and Zisserman, 2014) | – | ✓ | 43.2 | 45.9 | 42.6 | 44.4 | 47.1 | 49.0 |
| LFCC-LCNN (Monteiro et al., 2019) | – | ✓ | 44.3 | 47.8 | 47.4 | 50.3 | 48.5 | 50.1 |
| RawNet (Jung et al., 2019) | – | ✓ | 54.1 | 56.2 | 46.8 | 51.2 | 56.7 | 59.3 |
| ECAPA-TDNN (Desplanques et al., 2020) | – | ✓ | 67.3 | 69.8 | 59.8 | 62.7 | 70.2 | 72.6 |
| AASIST (Jung et al., 2022) | – | ✓ | 64.8 | 68.4 | 53.4 | 55.1 | 67.5 | 69.8 |
| ***ART-ADF*** | | ✓ | 67.0 | 69.7 | 66.3 | 68.9 | 70.2 | 73.7 |
| EmotionForensics (Mittal et al., 2020) | ✓ | ✓ | 80.6 | 84.4 | 78.1 | 79.8 | 88.7 | 91.9 |
| MDS (Chugh et al., 2020) | ✓ | ✓ | 89.8 | 91.6 | 82.8 | 86.5 | 92.0 | 94.3 |
| BA-TFD (Cai et al., 2022) | ✓ | ✓ | 79.1 | 84.6 | 80.8 | 84.9 | 92.1 | 94.9 |
| VFD (Cheng et al., 2023) | ✓ | ✓ | 80.9 | 85.1 | 81.5 | 86.1 | 93.4 | 95.6 |
| AVFakeNet (Ilyas et al., 2023) | ✓ | ✓ | 82.8 | 86.2 | 78.4 | 83.4 | 91.8 | 93.7 |
| AVoiD-DF (Yang et al., 2023) | ✓ | ✓ | 91.4 | 94.8 | 83.7 | 89.2 | 95.3 | 97.6 |
| PVASS-MDD (Yu et al., 2023) | ✓ | ✓ | **96.3** | **98.9** | 95.7 | 97.3 | **97.7** | **99.1** |
| ***ART-AVDF*** | ✓ | ✓ | 93.8 | 97.1 | **96.4** | **98.2** | 96.8 | 98.7 |

**Table 3**
The AUCs of cross-dataset for generalization ability evaluation. The training datasets are shown in the first row, and the testing sets are shown in the second row.

| Model | DFDC | | FakeAVCele | | DefakeAVMiT | |
|---|---|---|---|---|---|---|
| | FakeAVCele | DefakeAVMiT | DFDC | DefakeAVMiT | DFDC | FakeAVCele |
| EmotionForensics (Mittal et al., 2020) | 71.6 | 74.4 | 67.2 | 70.6 | 79.1 | 78.5 |
| MDS (Chugh et al., 2020) | 72.7 | 76.8 | 73.1 | 75.2 | 81.6 | 80.8 |
| VFD (Cheng et al., 2023) | 76.1 | 78.3 | 75.5 | 77.2 | 84.1 | 79.2 |
| AVFakeNet (Ilyas et al., 2023) | 70.1 | 73.1 | 69.6 | 71.6 | 79.1 | 77.2 |
| AVoiD-DF (Yang et al., 2023) | 82.8 | 84.4 | 80.7 | 83.2 | 90.3 | 88.7 |
| PVASS-MDD (Yu et al., 2023) | **87.7** | 88.9 | 84.8 | 87.5 | **95.3** | **93.6** |
| ***ART-AVDF*** | 88.5 | **91.4** | **86.7** | **89.5** | 93.2 | 91.6 |

**Table 4**
Ablation study for the effectiveness of each module on three benchmarks.

| Ablation model | DFDC | | FakeAVCele | | DefakeAVMiT | |
|---|---|---|---|---|---|---|
| | ACC | AUC | ACC | AUC | ACC | AUC |
| Visual only | 81.2 | 85.4 | 80.7 | 84.3 | 80.6 | 82.7 |
| Audio only | 67.0 | 69.7 | 66.3 | 68.9 | 70.2 | 73.7 |
| Articulatory predictor (AAI (Wu et al., 2023a)) | 86.9 | 87.3 | 90.4 | 93.9 | 91.1 | 93.8 |
| Lip encoder (Martinez et al., 2020) | 88.2 | 91.8 | 90.6 | 91.5 | 89.3 | 92.2 |
| Lip encoder (Ma et al., 2021) | 91.7 | 93.7 | 90.5 | 92.7 | 92.5 | 95.0 |
| Lip encoder (Ma et al., 2022) | 93.6 | 95.2 | 94.1 | 96.8 | 92.4 | 95.3 |
| w/o articulatory predictor | 77.6 | 79.8 | 81.2 | 85.3 | 82.0 | 85.7 |
| w/o lip encoder | 80.3 | 83.3 | 81.6 | 84.2 | 82.9 | 84.4 |
| w/o ART | 69.3 | 71.0 | 72.6 | 76.7 | 72.4 | 74.6 |
| w/o $\mathcal{L}_{ce}$ | 92.7 | 95.4 | 92.3 | 95.6 | 90.1 | 92.5 |
| w/o $\mathcal{L}_{con}$ | 89.9 | 92.1 | 91.5 | 94.0 | 92.8 | 94.6 |
| ***ART-AVDF*** | **93.8** | **97.1** | **96.4** | **98.2** | **96.8** | **98.7** |

an improvement of 0.7% in ACC and 0.9% in AUC on FakeAVCeleb, while underperforming on DFDC and DefakeAVMiT, with decreases of 1.7% (ACC) and 1.1% (AUC). This may be due to our pre-trained *ART* module failing to capture lip motion accurately, especially when the head poses involve large yaw and roll angles.

### 4.3. Generalization test

Since deepfake techniques applied for DFDC, FakeAVCeleb, and DefakeAVMiT are different, it is crucial for detection models to enhance their generalization ability across different datasets. Following the experiment setup outline in Yang et al. (2023) and Yu et al. (2023), we train our *ART-AVDF* on each benchmark and test on the other two datasets.

Table 3 illustrates the cross-dataset results of audio–visual deepfake models. Overall, we observe a drop in AUC for all models when detecting unseen videos. Regarding the comparison of models trained on the DFDC and FakeAVCeleb dataset, our *ART-AVDF* outperforms all other models, achieving 88.5% and 91.4%, and 86.7% and 89.5% on the other two datasets respectively. For models trained on DefakeAVMiT dataset, we observe that our *ART-AVDF* receives a slight improvement than AVoid-DF and PVASS-MMD. Such results demonstrate that our framework has better generalization ability.

### 4.4. Ablation study

To perform extensive ablation experiments, we train our *ART-AVDF* under different settings and discuss the effectiveness of each component. The results are shown in Table 4.

(1) *Effectiveness of Audio–Visual Joint Learning.* As discussed in , both auditory and visual data may be manipulated. In this paper, we propose a plug-in module, *ART*, to provide articulatory representations that capture both articulation movement and lip movement. When conducting the unimodal ablation study, we remove the auditory (or visual) branch to perform deepfake detection. Overall, uni-modal learning of visual or audio gains poor performances on three benchmarks, averaging — visual only: 80.8% (ACC) and 84.1% (AUC); audio only: 67.8% (ACC) and 70.8% (AUC). On the other hand, the multi-modal model achieves promising results.

(2) *Effectiveness of ART Module.* While good performances are observed by merely combining audio and visual information, we aim to further explore the audio–visual correspondence in terms of the physiological information implicitly involved during human speech. To evaluate the effectiveness of the proposed *ART*, seven different training strategies have been analyzed, including AAI as the articulatory predictor, different lip encoders, and without different components. As shown in Table 4, we can observe that the model without *ART* module obtains the lowest results, due to the absence of comprehensive audio–visual information and the simple joint audio–visual learning could hinder the performance of detectors. When utilizing the AAI model (Wu et al., 2023a) as the articulatory predictor and other lip encoders (Martinez et al., 2020; Ma et al., 2021, 2022), both ACC and AUC results decrease for all datasets. Moreover, we stress that the performance of articulatory representation learning is not our ultimate objective. *ART* module is designed to be a proxy for learning rich audio–visual representations.

(3) *Effectiveness of loss functions in AVDF Module.* Results in Table 4 indicate the importance of the cross-modal loss function, which takes the weighted sum of Cross-Entropy Loss ($\mathcal{L}_{ce}$) and Contrastive Loss ($\mathcal{L}_{con}$).

### 4.5. Qualitative evaluation

In this section, we present qualitative results of *ART-AVDF*.

(1) *Mass Results.* As shown in Fig. 6(a), we demonstrate detection results of the recently released dataset, GOTCHA (Mittal et al., 2022), which consists of 49,603 fake videos created by DeepFaceLab, FSGANv2, and LIA, and 816 real videos. This dataset covers different scenarios, *e.g.*, side flashing, facial expressions, and facial occlusion with hands, sunglasses, and clothes. Overall, our *ART-AVDF* achieves 89.2% of ACC and 91.0% of AUC. Moreover, we also evaluate our *ART-AVDF* on YouTube videos with varying lighting conditions and diverse backgrounds (shown in Fig. 6(b)).

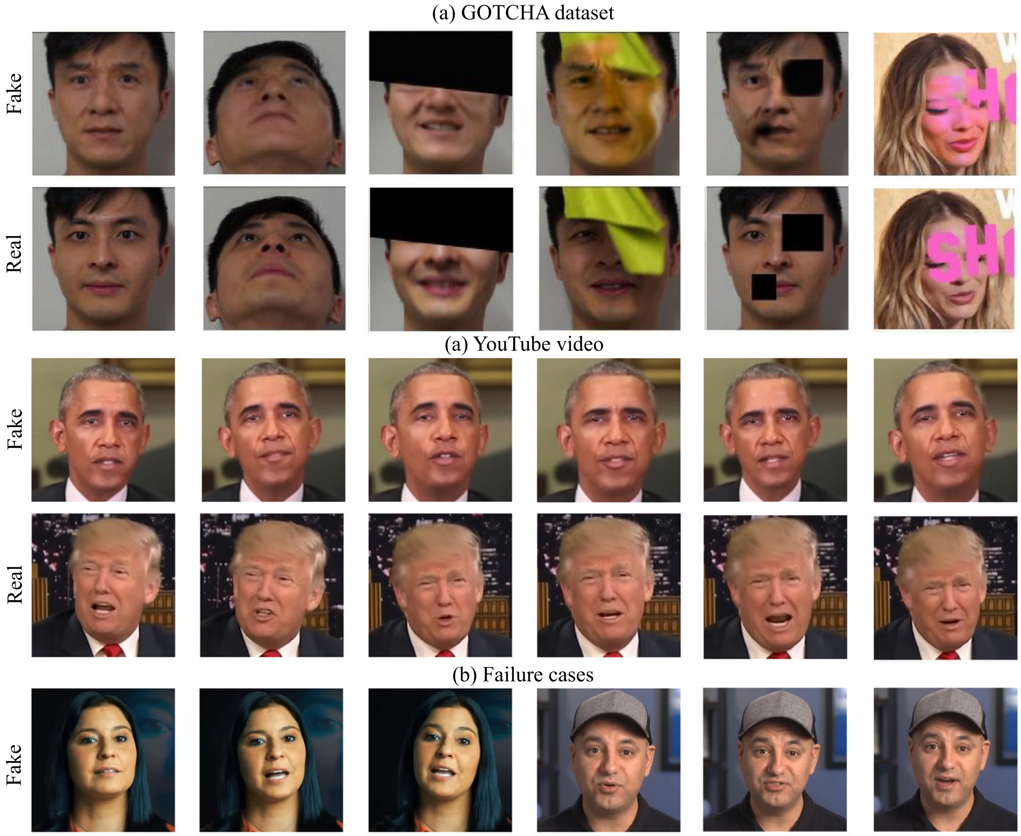


**Fig. 6.** Qualitative results of our proposed *ART-AVDF*: (a) accurately detected results of GOTCHA dataset; (b) accurately detected results of YouTube videos; and (c) failure cases, *i.e.* fake videos classified as real ones.

(2) *Failure Cases.* Our *ART-AVDF* leverages the audio–visual articulatory representations to further model the correlation between auditory and visual information. However, the *ART* module could limit the detection performance in some cases — our method may classify a fake video as a real video due to well-rendered face textures, as shown in Fig. 6(c). This is because humans exhibit significant variations in unique facial expressions and gestures, accompanying spoken words influenced by distinct accents, speed, and intonations. Such variations complicate the process of analyzing and interpreting audio–visual data, highlighting the need for algorithms capable of handling these diverse speech patterns. Moreover, our *ART-AVDF* may experience reduced accuracy when the lip is occluded or cannot be detected due to large pitch and yaw angles in the head pose. However, our audio–visual articulatory representation has minimized these limitations as much as possible.

## 5. Conclusion

The field of deepfake detection is moving towards a multimodal approach that integrates data from both audio and visual channels to enhance detection accuracy. In this paper, we propose *ART-AVDF*, an audio–visual joint learning framework for multi-modal deepfake detection by using articulatory representations. Inspired by the substantial physiological expressions that occur during speech, we introduce an articulatory representation learning (*ART*) module to model the correlation between articulation movement and lip motion by utilizing the self-supervised learning strategy. The audio–visual deepfake detection (*AVDF*) module is then presented for jointly multi-modal learning with the audio–visual articulatory embedding. Experimental results demonstrate that *ART-AVDF* outperforms both unimodal and multi-modal detection models, indicating that articulatory inconsistencies between the two modalities could be an effective cue for deepfake detection.

The multi-modal deepfake detection remains an open problem and a challenging task. In future work, we will further improve the performance of the *ART* module and incorporate additional factors, *e.g.*, distinct accents and nuanced facial expressions, to create a robust model capable of addressing the rich diversity in human communication.

## CRediT authorship contribution statement

**Yujia Wang:** Writing – original draft, Validation, Methodology, Investigation, Conceptualization. **Hua Huang:** Supervision, Conceptualization.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgments

Yujia Wang was supported by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No. 62302046) when working on this project.

## References

Afchar, D., Nozick, V., Yamagishi, J., Echizen, I., 2018. Mesonet: a compact facial video forgery detection network. In: International Workshop on Information Forensics and Security. WIFS, IEEE, pp. 1–7.

Afouras, T., Owens, A., Chung, J.S., Zisserman, A., 2020. Self-supervised learning of audio-visual objects from video. In: ECCV. Springer, pp. 208–224.

Alayrac, J.-B., Recasens, A., Schneider, R., Arandjelović, R., Ramapuram, J., De Fauw, J., Smaira, L., Dieleman, S., Zisserman, A., 2020. Self-supervised multimodal versatile networks. Neural Inf. Process. Syst. 33, 25–37.

Bigioi, D., Basak, S., Stypułkowski, M., Zieba, M., Jordan, H., McDonnell, R., Corcoran, P., 2024. Speech driven video editing via an audio-conditioned diffusion model. Image Vis. Comput. 142, 104911.

Birkholz, P., 2013. Modeling consonant-vowel coarticulation for articulatory speech synthesis. PLoS One 8 (4), e60603.

Cai, Z., Ghosh, S., Dhall, A., Gedeon, T., Stefanov, K., Hayat, M., 2023. Glitch in the matrix: A large scale benchmark for content driven audio–visual forgery detection and localization. Comput. Vis. Image Underst. 236, 103818.

Cai, Z., Stefanov, K., Dhall, A., Hayat, M., 2022. Do you really mean that? Content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In: International Conference on Digital Image Computing: Techniques and Applications. DICTA, IEEE, pp. 1–10.

Cheng, H., Guo, Y., Wang, T., Li, Q., Chang, X., Nie, L., 2023. Voice-face homogeneity tells deepfake. ACM Trans. Multimed. 20 (3), 1–22.

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

Cozzolino, D., Pianese, A., Nießner, M., Verdoliva, L., 2023. Audio-visual person-of-interest deepfake detection. In: CVPR. pp. 943–952.

Desplanques, B., Thienpondt, J., Demuynck, K., 2020. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. arXiv preprint arXiv:2005.07143.

Diehl, R.L., 2008. Acoustic and auditory phonetics: the adaptive design of speech sound systems. Phil. Trans. R. Soc. B 363, 965–978.

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C., 2020. The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397.

Eshky, A., Ribeiro, M.S., Cleland, J., Richmond, K., Roxburgh, Z., Scobbie, J., Wrench, A., 2019. UltraSuite: a repository of ultrasound and acoustic data from child speech therapy sessions. arXiv preprint arXiv:1907.00835.

Fang, Q., Dang, J., 2006. Speech synthesis based on a physiological articulatory model. In: International Symposium on Chinese Spoken Language Processing. Springer, pp. 211–222.

Fant, G., 1960. Acoustic Theory of Speech Production.

Feng, C., Chen, Z., Owens, A., 2023. Self-supervised video forensics by audio-visual anomaly detection. In: CVPR. pp. 10491–10503.

Fujisaki, H., 1983. Dynamic characteristics of voice fundamental frequency in speech and singing. Prod. Speech 39–55.

Gaines, J.L., shik Kim, K., Parrell, B., Ramanarayanan, V., Nagarajan, S.S., Houde, J.F., 2021. Discrete constriction locations describe a comprehensive range of vocal tract shapes in the Maeda model. Jasa Express Lett. 1.

Gérard, J.-M., Wilhelms-Tricarico, R., Perrier, P., Payan, Y., 2006. A 3D dynamical biomechanical tongue model to study speech motor control. arXiv preprint physics/0606148.

Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M., 2021. Lips don't lie: A generalisable and robust approach to face forgery detection. In: CVPR. pp. 5039–5049.

Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al., 2017. CNN architectures for large-scale audio classification. In: ICASSP. IEEE, pp. 131–135.

Hershey, J., Movellan, J., 1999. Audio vision: Using audio-visual synchrony to locate sounds. Neural Inf. Process. Syst. 12.

Hu, J., Liao, X., Liang, J., Zhou, W., Qin, Z., 2022. Finfer: Frame inference-based deepfake detection for high-visual-quality videos. In: AAAI. pp. 951–959.

Huang, X., Belongie, S., 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In: ICCV. pp. 1501–1510.

Ilyas, H., Javed, A., Malik, K.M., 2023. AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio–visual deepfakes detection. Appl. Soft Comput. 136, 110124.

Jung, J.-w., Heo, H.-S., Kim, J.-h., Shim, H.-j., Yu, H.-J., 2019. Rawnet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification. arXiv preprint arXiv:1904.08104.

Jung, J.-w., Heo, H.-S., Tak, H., Shim, H.-j., Chung, J.S., Lee, B.-J., Yu, H.-J., Evans, N., 2022. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In: ICASSP. IEEE, pp. 6367–6371.

Kempelen, W.v., Brekle, H.E., Wildgen, W., 1791. Mechanismus der menschlichen Sprache nebst Beschreibung einer sprechenden Maschine. Gramm. Universalis 4.

Khalid, H., Kim, M., Tariq, S., Woo, S.S., 2021a. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In: Workshop on Synthetic Multimedia-Audiovisual Deepfake Generation and Detection. pp. 7–15.

Khalid, H., Tariq, S., Kim, M., Woo, S.S., 2021b. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. arXiv preprint arXiv:2108.05080.

Korbar, B., Tran, D., Torresani, L., 2018. Cooperative learning of audio and video models from self-supervised synchronization. Adv. Neural Inf. Process. Syst. 31.

Kum, S., Nam, J., 2019. Joint detection and classification of singing voice melody using convolutional recurrent neural networks. Appl. Sci. 9 (7), 1324.

Lee, J., Shaiman, S., Weismer, G., 2016. Relationship between tongue positions and formant frequencies in female speakers. J. Acoust. Soc. Am. 139 1, 426–440.

Lin, Q., Fant, G., 1989. Vocal-tract area-function parameters from formant frequencies. In: EUROSPEECH.

Lindblom, B.E., Sundberg, J.E., 1971. Acoustical consequences of lip, tongue, jaw, and larynx movement. J. Acoust. Soc. Am. 50 (4B), 1166–1179.

Ling, Z.-H., Richmond, K., Yamagishi, J., Wang, R.-H., 2009. Integrating articulatory features into HMM-based parametric speech synthesis. IEEE Trans. Audio Speech Lang. Process. 17 (6), 1171–1185.

Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H., 2022. Video swin transformer. In: CVPR. pp. 3202–3211.

Liy, C.M., InIctuOculi, L., 2018. Exposing AI created fake videos by detecting eye blinking. In: International Workshop on Information Forensics and Security. WIFS, IEEE.

Ma, P., Martinez, B., Petridis, S., Pantic, M., 2021. Towards practical lipreading with distilled and efficient models. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 7608–7612.

Ma, P., Wang, Y., Petridis, S., Shen, J., Pantic, M., 2022. Training strategies for improved lip-reading. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 8472–8476.

Malik, H., Farid, H., 2010. Audio forensics from acoustic reverberation. In: ICASSP. IEEE, pp. 1710–1713.

Martinez, B., Ma, P., Petridis, S., Pantic, M., 2020. Lipreading using temporal convolutional networks. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 6319–6323.

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D., 2020. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In: ACM MM. pp. 2823–2832.

Mittal, G., Yenphraphai, J., Hegde, C., Memon, N., 2022. Gotcha: A challenge-response system for real-time deepfake detection. arXiv preprint arXiv:2210.06186.

Monteiro, J., Alam, J., Falk, T.H., 2019. End-to-end detection of attacks to automatic speaker recognizers with time-attentive light convolutional neural networks. In: International Workshop on Machine Learning for Signal Processing. MLSP, IEEE, pp. 1–6.

Morgado, P., Vasconcelos, N., Misra, I., 2021. Audio-visual instance discrimination with cross-modal agreement. In: CVPR. pp. 12475–12486.

Motie-Shirazi, M., Zañartu, M., Peterson, S.D., Erath, B.D., 2021. Vocal fold dynamics in a synthetic self-oscillating model: Intraglottal aerodynamic pressure and energy. J. Acoust. Soc. Am. 150 (2), 1332–1345.

Muppalla, S., Jia, S., Lyu, S., 2023. Integrating audio-visual features for multimodal deepfake detection. arXiv preprint arXiv:2310.03827.

Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., Nayak, K., Kim, Y.-C., Zhu, Y., Goldstein, L., et al., 2014. Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). J. Acoust. Soc. Am. 136 (3), 1307–1311.

Nguyen, T.T., Nguyen, Q.V.H., Nguyen, D.T., Nguyen, D.T., Huynh-The, T., Nahavandi, S., Nguyen, T.T., Pham, Q.-V., Nguyen, C.M., 2022. Deep learning for deepfakes creation and detection: A survey. Comput. Vis. Image Underst. 223, 103525.

Nguyen, H.H., Yamagishi, J., Echizen, I., 2019. Capsule-forensics: Using capsule networks to detect forged images and videos. In: ICASSP. IEEE, pp. 2307–2311.

Oh, T.-H., Dekel, T., Kim, C., Mosseri, I., Freeman, W.T., Rubinstein, M., Matusik, W., 2019. Speech2face: Learning the face behind a voice. In: CVPR. pp. 7539–7548.

Perkell, J.S., 1974. A physiologically-oriented model of tongue activity in speech production.

Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J., 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: ECCV. Springer, pp. 86–103.

Raza, M.A., Malik, K.M., 2023. Multimodaltrace: Deepfake detection using audiovisual representation learning. In: CVPR. pp. 993–1000.

Richmond, K., Hoole, P., King, S., 2011. Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In: International Speech Communication Association.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M., 2019. Faceforensics++: Learning to detect manipulated facial images. In: ICCV. pp. 1–11.

Scherer, K., 1992. What does a facial expression express.

Seneviratne, N., Sivaraman, G., Espy-Wilson, C.Y., 2019. Multi-corpus acoustic-to-articulatory speech inversion. In: Interspeech. pp. 859–863.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Tiede, M., Espy-Wilson, C.Y., Goldenberg, D., Mitra, V., Nam, H., Sivaraman, G., 2017. Quantifying kinematic aspects of reduction in a contrasting rate production task. J. Acoust. Soc. Am. 141 (5_Supplement), 3580.

Toda, T., Black, A.W., Tokuda, K., 2004. Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis. In: ISCA Workshop on Speech Synthesis.

Wilhelms-Tricarico, R., 1995. Physiological modeling of speech production: Methods for modeling soft-tissue articulators. J. Acoust. Soc. Am. 97 (5), 3085–3098.

Wu, P., Chen, L.-W., Cho, C.J., Watanabe, S., Goldstein, L., Black, A.W., Anumanchipalli, G.K., 2023a. Speaker-independent acoustic-to-articulatory speech inversion. In: ICASSP. IEEE, pp. 1–5.

Wu, P., Li, T., Lu, Y., Zhang, Y., Lian, J., Black, A.W., Goldstein, L., Watanabe, S., Anumanchipalli, G.K., 2023b. Deep speech synthesis from MRI-based articulatory representations. arXiv preprint arXiv:2307.02471.

Yang, X., Li, Y., Lyu, S., 2019. Exposing deep fakes using inconsistent head poses. In: ICASSP. IEEE, pp. 8261–8265.

Yang, W., Zhou, X., Chen, Z., Guo, B., Ba, Z., Xia, Z., Cao, X., Ren, K., 2023. Avoid-df: Audio-visual joint learning for detecting deepfake. Trans. Inf. Forensics Secur. 18, 2015–2029.

Yi, J., Fu, R., Tao, J., Nie, S., Ma, H., Wang, C., Wang, T., Tian, Z., Bai, Y., Fan, C., et al., 2022. Add 2022: the first audio deep synthesis detection challenge. In: ICASSP. IEEE, pp. 9216–9220.

Yi, J., Wang, C., Tao, J., Zhang, X., Zhang, C.Y., Zhao, Y., 2023. Audio deepfake detection: A survey. arXiv preprint arXiv:2308.14970.

Yu, Y., Liu, X., Ni, R., Yang, S., Zhao, Y., Kot, A.C., 2023. Pvass-mdd: predictive visual-audio alignment self-supervision for multimodal deepfake detection. Trans. Circuits Syst. Video Technol..

Yu, Y., Shandiz, A.H., Tóth, L., 2021. Reconstructing speech from real-time articulatory MRI using neural vocoders. In: European Signal Processing Conference. EUSIPCO, IEEE, pp. 945–949.

Zhang, Y., Lin, W., Xu, J., 2024. Joint audio-visual attention with contrastive learning for more general deepfake detection. ACM Trans. Multimed. Comput. Commun. Appl. 20 (5), 1–23.

Zhang, J.-X., Richmond, K., Ling, Z.-H., Dai, L., 2021. Talnet: Voice reconstruction from tongue and lip articulation with transfer learning from text-to-speech synthesis. In: AAAI. pp. 14402–14410.

Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., Yu, N., 2021. Multi-attentional deepfake detection. In: CVPR. pp. 2185–2194.

Zhao, H., Zhou, W., Chen, D., Zhang, W., Yu, N., 2022. Self-supervised transformer for deepfake detection. arXiv preprint arXiv:2203.01265.

Zhou, Y., Lim, S.-N., 2021. Joint audio-visual deepfake detection. In: CVPR. pp. 14800–14809.

Zong, Y., Mac Aodha, O., Hospedales, T., 2023. Self-supervised multimodal learning: A survey. arXiv preprint arXiv:2304.01008.

Zou, H., Shen, M., Hu, Y., Chen, C., Chng, E.S., Rajan, D., 2024. Cross-modality and within-modality regularization for audio-visual deepfake detection. arXiv preprint arXiv:2401.05746.