SpeechForensics: Audio-Visual Speech Representation Learning for Face Forgery Detection

Yachao Liang[1,2] **Min Yu**[1,2*] **Gang Li**[3] **Jianguo Jiang**[1,2] **Boquan Li**[4*]
Feng Yu[5] **Ning Zhang**[6] **Xiang Meng**[1,2] **Weiqing Huang**[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]Deakin University [4]Harbin Engineering University
[5]Institute of Computing Technology, Chinese Academy of Sciences
[6]Institute of Forensic Science, Ministry of Public Security
{liangyachao, yumin}@iie.ac.cn

Abstract

Detection of face forgery videos remains a formidable challenge in the field of digital forensics, especially the generalization to unseen datasets and common perturbations. In this paper, we tackle this issue by leveraging the synergy between audio and visual speech elements, embarking on a novel approach through audio-visual speech representation learning. Our work is motivated by the finding that audio signals, enriched with speech content, can provide precise information effectively reflecting facial movements. To this end, we first learn precise audio-visual speech representations on real videos via a self-supervised masked prediction task, which encodes both local and global semantic information simultaneously. Then, the derived model is directly transferred to the forgery detection task. Extensive experiments demonstrate that our method outperforms the state-of-the-art methods in terms of cross-dataset generalization and robustness, without the participation of any fake video in model training. The code is available here.

1 Introduction

The rapid advancement of generative models enables synthetic realistic facial images [20, 39, 35], and they have significantly enhanced face manipulation techniques, allowing for the replacement of facial identities and the modification of attributes such as expressions [60, 59] and lip movements[1, 3]. While these advancements offer vast potential for entertainment and filmmaking, they also harbor the risk of misuse for deceptive purposes.

In response to these concerns, there has been a surge in the development of face forgery detection methodologies grounded in deep learning [4, 55, 63, 71, 43, 70, 30, 64]. Despite these efforts, it is widely acknowledged that face forgery detectors frequently experience a decline in effectiveness when confronted with novel manipulation techniques [19, 12, 45]. This vulnerability poses significant hurdles to the reliable application of these detection systems, highlighting a critical area for ongoing research and innovation.

To enhance the generalization capabilities of face forgery detectors, researchers have proposed various methods aimed at mining more discriminative clues [70, 62, 71, 29, 64]. Some works focus on detecting spatial artifacts left in the process of facial manipulation [12, 43, 70, 9], especially blending boundaries [43, 57]. However, these methods are sensitive to common perturbations, making them difficult to generalize to real-life scenarios. Another line of research resorts to model

[*] Corresponding authors

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

temporal features [67, 29, 69, 64], considering that fake videos are synthesized in a frame-by-frame manner. They identify unnatural facial movements existing in fake videos by applying special architectures [71] or introducing auxiliary tasks [18, 30, 29]. Although showing promising results, short-term information modeling capacity (e.g., 1 second [71, 30]) makes them overfit to specific low-level temporal features to varying degrees, resulting in their suboptimal generalization on unseen datasets and perturbations, as observed in our experiments.

This motivates us to find more general semantic-level features to detect anomalous facial movements. Recent efforts on audio-visual speech recognition have shown that accurate speech contents can be extracted from both audio signals and lip movements [56, 28, 46]. Inspired by this, we conjecture that audio signals could provide strong semantic supervision for identifying inaccurate lip movements in fake videos, given that lip sequences and audio segments in a real video should convey the same speech contents. This brings us to the key problem: *how to extract semantically rich speech-related features to represent detailed lip movements?*

An intuitive solution is to align the speech representations of each frame of audio segments and lip sequences directly, as in [17]. However, this method will fail to detect fake videos processed by lip synchronization techniques, such as `Wav2Lip` [53], commonly used by recent talking face generation technologies. Considering that local lip synchronization cannot bring long-range temporal coherence, we further propose to perform forgery detection by learning audio and visual speech representations in a framework that encodes both phonetic and linguistic information, which we term as local and global information. Specifically, it learns local information by frame-wisely audio-visual representation alignment and models global dependencies via masked prediction task, following previous speech representation learning methods [56, 28, 73]. In this way, both short-range and long-range temporal features are learned. After learning audio-visual speech representation on real videos, we directly transfer the trained model to the forgery detection task by finding discrepancies between visual and audio speech representations in fake videos.

Thanks to the unsupervised manner and high-level semantic learning, our method, termed `SpeechForensics`, avoids overfitting on forgery features and shows strong robustness on various perturbations. We conduct comprehensive experiments to evaluate the effectiveness of our method, and it shows strong performance under different manipulations, datasets, and perturbations. Especially, our method achieves the AUC of 99.0% on FakeAVCeleb [38] and 91.7% on KoDF [41]. Our main contributions are summarized as follows:

- We propose to perform face forgery video detection by extracting speech representations from audio and visual streams. It learns on real videos and can smoothly transfer to the forgery detection task, markedly streamlining the forgery detection workflow.
- We demonstrate a simple framework, which encodes both short-range and long-range temporal information, is well-suited to our method. And we tailor it to a face forgery detector using the proposed modality alignment module.
- Extensive experiments demonstrate the superiority of our method over the state-of-the-art methods in terms of cross-dataset generalization, robustness, and interpretability, in an unsupervised manner.

2 Related Work

Face Forgery Detection. Initial approaches in face forgery detection predominantly treat the task as a binary classification problem, leveraging deep learning models trained on datasets specifically compiled for detecting forgeries [4, 55, 21, 42]. For instance, [4] introduces a pair of detection networks known as `Mesonet` and `MesoInception`, demonstrating that lightweight neural networks can effectively undertake forgery detection tasks. Analogously, [55] highlights the superior performance of an unconstrained `Xception` network over its predecessors, focusing primarily on the analysis of spatial details within individual frames. Subsequently, some works [27, 8, 68, 26, 22, 25] try to combine temporal networks to perform forgery detection. Despite the promising results in the in-dataset setting, these vanilla methods usually suffer from severe performance degradation when facing unseen forgeries.

General Forgery Detection. To boost the generalization of detectors on unseen forgeries, researchers attempt to find more discriminative features at both image and video levels.

Image-based methods [12, 43, 70, 57, 9] analyze the spatial artifacts common to forged faces and generate synthetic data to guide models to focus on them. For example, `Face X-ray` [43] and `SBI` [57] detect blending boundaries caused by the fusion of the forged face and background, and `AUNet` [9] concentrates on the relation between different facial action units. While they are adept at identifying specific artifacts, these artifacts are easily destroyed by some common perturbations, e.g., compression, which makes it difficult to generalize to real scenarios.

On another front, video-based methods make efforts to explore temporal clues via special network architectures [71, 64] or auxiliary tasks [44, 30, 29, 69]. `FTCN` [71] reduces the convolutional kernel size to 1 forcing the network to only focuses on temporal features. `LipForensics` [30] uncovers unnatural lip movements via pre-training on the lipreading task and finetuning on forgery datasets. Analogously, `RealForensics` [29] leverages cross-modal self-supervision learning to capture facial movements. Despite their notable performance, they tend to rely on short-range low-level temporal features, leading to their limited generalization on new datasets and robustness against common perturbations. In contrast, our method detects both short-range and long-range anomalous facial movements using semantic-level information and functions in an unsupervised manner, inherently possessing superior generalization and robustness.

Audio-Visual Speech Representation Learning. The advent of extensive large-scale audio-visual speech datasets [6, 7, 16] has spurred the development of numerous audio-visual speech representation learning methods in recent years [17, 53, 56, 46]. [17] learns visual and audio speech representations simultaneously based on synchronization signals between lip movements and corresponding audio segments. Benefiting from off-the-shelf audio speech recognition models, [47] proposes to learn visual speech representations by minimizing the distance between learned visual embeddings and pre-trained audio embeddings. [28] further advances this field by simultaneously learning visual and auditory speech representations through student-teacher networks. We choose `AVHuBERT` [56] as the implementation of our audio-visual speech representation learning, considering it fits into our framework, i.e., learning both local and global semantic representations, and demonstrates remarkable efficacy in downstream speech recognition tasks.

3 Method

Our method consists of the audio-visual speech representation learning stage and the face forgery detection stage. We first learn semantically rich visual and audio speech representations in a unified feature space from real videos, which can be implemented by many audio-visual speech representation learning approaches [56, 73, 28]. Subsequently, the model leverages these representations to pinpoint discrepancies between lip movements and corresponding audio segments in fake videos.

3.1 Speech Representation Learning

In order to simultaneously learn local and global speech information, strongly correlated with short-range and long-range lip movements, we conduct frame-wise audio-visual representation alignment and the masked prediction task. Since we mainly focus on the forgery detection task, we will briefly introduce the representation learning stage, for details refer to [56].

Local representation alignment. Considering $\mathcal{X} = \{(\boldsymbol{I^i}, \boldsymbol{A^i})\}_{i=1}^{N}$ as the set of audio-visual pairs extracted from real videos. Given a visual and audio pair $(\boldsymbol{I}_{1:T}, \boldsymbol{A}_{1:T})$ from the set $\mathcal{X}$, where T represents the sequence length of clip. We first get intermediate features $F^v_{1:T} = f^v_e(\boldsymbol{I}_{1:T})$ and $F^a_{1:T} = f^a_e(\boldsymbol{A}_{1:T})$ through the visual frontend f^v_e and the audio frontend f^a_e, respectively. And F^v and F^a are fused by channel-wise concatenation, before which modality dropout is applied to allow the unimodal input. Then the fused features are fed into subsequent transformer encoder to learn frame-wise speech representations, as shown in fig. 1. The target labels for training are derived through cluster assignment [32], which are initialized based on `MFCC` features of audio and iteratively refined with audio-visual features learned by encoders via k-means. And we denote them as $\gamma_{1:T} \in \{1, 2, \ldots, C\}$, where C is the size of the codebook. By this means, the representations of every frames of visual and audio modalities are aligned in a unified feature space.

Global information modeling. Following the described procedure, the visual and audio speech representations for each frame are synchronized, facilitating the computation of their similarity in subsequent analyses. However, local speech contents conveyed by lip movements, i.e., visemes, only

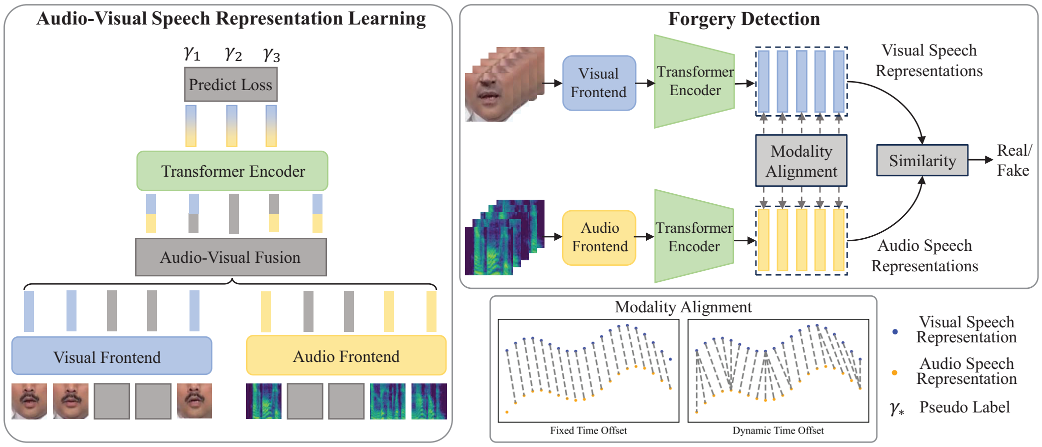


Figure 1: **Overview of the proposed method.** During the stage of audio-visual speech representation learning, local speech representations and global information are learned by frame-wise feature alignment and the masked prediction task, respectively. In the stage of forgery detection, we separately feed the whole lip movement sequence and audio stream of a video into the learned model to get visual and audio speech representations. And we flag videos with low matching scores between visual and audio speech representations as fake videos.

contain limited temporal features and can easily be tampered by lip-sync methods, e.g., `Wav2Lip` [53]. To address this problem, we further introduce global temporal information modeling.

We employ the masked-prediction task, a method extensively utilized across various fields [37, 36, 11], to model contextual dependencies effectively. Let $M_v, M_a \subset \{1, 2, \dots, T\}$ represent the sets of indices of visual and audio masked sequences, this task can be formulated as:

$$\mathcal{L} = \sum_{t \in M_v \cup M_a} \log p(\gamma_t \mid \tilde{F}^v, \tilde{F}^a) \tag{1}$$

where $\tilde{F}^a$ and $\tilde{F}^v$ denotes of the corrupted audio features and visual features, respectively.

3.2 Face Forgery Detection

For the forgery detection, we aim to identify discrepancies between visual and audio speech representations in manipulated videos. To achieve this, we extract visual embeddings as visual speech representations by feeding the learned model with only visual inputs, and apply the same procedure for audio embeddings. Notably, obtaining accurate speech representations from the audio stream is more straightforward, allowing these to function effectively as pseudo-labels.

Given the visual and audio speech representations of any video, we get their matching score by calculating frame-wise cosine similarity, which can be formulated as:

$$\mathcal{S}(e^v, e^a) = \frac{1}{T} \sum_{t=1}^{T} sim(e^v_t, e^a_t) \tag{2}$$

where e_v and e_a represent the visual and audio embeddings extracted from the final layer of our model, respectively, and $sim(\cdot, \cdot)$ is the cosine similarity between two vectors. And videos exhibiting low matching scores are consequently classified as forgeries.

Another crucial problem is the time offsets between visual and audio signals, which inevitably exist even in real videos due to recording or encoding errors [5, 24]. To address this problem, we consider two time offset assumptions, i.e., fixed and dynamic, and introduce different modality alignment approaches to alleviate their impact. See the intuitive illustration of these two assumptions in fig. 1.

Assumption 1: Fixed time offset. First, we assume that the frame offsets between visual and audio streams remain constant throughout the video. We apply a sliding-window technique, as in [17], to correct time offsets prior to calculating final matching scores. Specifically, based on the assumption that the maximum offset between visual and audio streams is τ, we compute the cosine similarity between the feature of each visual frame and its adjacent audio features within a window of $\pm\tau$

frames. Thereafter, the highest average cosine similarity across this window is taken as the overall similarity of the video. Consequently, the similarity eq. (2) can be re-formulated as:

$$\mathcal{S}(e^v, e^a) = \max_{t-\tau \leq i \leq t+\tau} \frac{1}{T} \sum_{t=1}^{T} sim(e_t^v, e_i^a) \tag{3}$$

where we pad zero vectors as e_i^a when $i < 1$ or $i > T$.

Assumption 2: Dynamic time offset. On the contrary, we also consider the dynamic assumption, i.e., the frame offsets between visual and audio signals vary over time. Based on this assumption, we introduce the *Dynamic Time Warping* (DTW) algorithm [49], commonly used to align and calculate the similarity of two time series data. In this case, eq. (2) will be re-written as:

$$\mathcal{S}(e^v, e^a) = DTW(e_{1:T}^v, e_{1:T}^a) \tag{4}$$

where we also apply cosine similarity as the cost measure of DTW.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate our methods across three distinct video forgery datasets: `Faceforensics++` (FF++) [55], `FakeAVCeleb` [38] and `KoDF` [41]. Note that Only the FakeAVCeleb contains videos belong to the Real-Visual-Fake-Audio category, and we exclude them as we focus on the facial forgery and to maintain fairness of experiments.

Faceforensics++ contains 1,000 real videos alongside 4,000 fake videos, created via four different manipulation methods. These include face swapping methods (`DeepFakes` [1], `FaceSwap` [3]), and two face reenactment methods (`Face2Face` [60] and `NeuralTextures` [59]). For our evaluation, We re-download videos using the provided YouTube IDs and extract audio segments from the provided frame locations. Contrary to the common practice of treating `Faceforensics++` as a visual-only dataset, we paired the original videos with their corresponding audio segments to create an audio-visual test dataset. After excluding videos unavailable or containing non-corresponding mouth movements and voices, we selected 500 videos from each category for testing.

FakeAVCeleb contains 500 real videos and 19,500 fake videos. It is derived from `VoxCeleb2` [16] and represents diverse ethnic backgrounds, ages and genders. This dataset involves four manipulation techniques, `Faceswap` [40] and `Faceswap GAN` (`FSGAN`) [50] for face swapping, `SV2TTS` [33] for real-time cloning voice, and `Wav2Lip` [53] for audio-driven facial reenactment.

KoDF [41] is a large-scale Korean forgery datasets, containing 62,166 real videos and 175,776 fake videos. For our evaluation, we focus on fake videos crafted using four distinct manipulation techniques: `FaceSwap` [3], `DeepFaceLab` [52], `FOMM` [58], `Audio-driven` (including ATFHP [66] and `Wav2Lip` [53]). From each of these categories, we randomly select 1,000 videos to compile our testing set.

Preprocessing. We first utilize FFmpeg [61] to convert all videos into 25fps and audio into 16kHz sample rate. For each video clip, we initiate the process by identifying faces using `RetinaFace` [23] and subsequently extract facial landmarks with `FAN` [10]. We then align the frames using affine transformations, and crop 96×96 regions centered around the mouth, as indicated by the landmarks. For the audio part, we extract `MFCC` features from wavform every 10ms, and we concatenate 4 adjacent audio frames to align with visual modality.

Architecture & Training. The audio-visual speech representation model consists of visual frontend, audio frontend and masked predictor. Following AVHuBert [56], We use the Resnet-18 2D+3D [48] as the visual frontend. And the audio frontend contains only a single linear projection layer to avoid the over-reliance issue on the audio stream [56]. The masked predictor is implemented by the standard transformer encoder. Further details can be found in appendix A.1.

The model is trained on `LRS3` [7] and `VoxCeleb2` [16] datasets, which contain 433 and 1326 hours of videos respectively. The training process adhere to the methodology outlined by [56]. For the purposes of our experiments, unless otherwise noted, we utilize a publicly available pretrained model [2].

[2] `https://github.com/facebookresearch/av_hubert`

4.2 Quantitative Comparisons

We compare our method with several state-of-the-art detectors, including Xception [55], Patch-based [12], Face X-ray [43], LipForensics [30], FTCN [71], RealForensics [29] and AVAD [13]. Moreover, we also construct a model, consisting of two Resnet 2D [31] models, to extract only phonetic-level (5 frames) speech information for comparison, And we dub it SpeechForensics-Local. See appendix A.2 for more details about above detectors.

We undertake extensive experiments focusing on the generalization and robustness of detectors. In line with established practices in the field [30, 29, 57, 9, 64], we utilize the area under the receiver operating characteristic curve (AUC) to gauge the efficacy of our method at the video level. Unless otherwise specified, we take the fixed time offset assumption.

Cross-Manipulation Generalization. In real-world situations, detectors frequently encounter novel manipulation techniques, underscoring the necessity for these systems to possess robust generalization capabilities against unseen manipulations. To assess our method's ability to generalize across different manipulation methods, we conducted evaluations on the widely recognized FF++ high-quality (HQ) dataset. For supervised methods, the experiments adopt the leave-one-out strategy, in line with established practices [30, 29, 71]. It is noteworthy that the settings for both cross-manipulation and cross-dataset are equivalent for unsupervised methods, i.e., AVAD [24] and our method.

The AUC results presented in table 1 demonstrate that our method either matches or exceeds performance across different categories, notably without utilizing any forgery samples. Remarkably, our approach achieves perfect results (100%) with the two face reenactment methods, Face2Face and NeuralTextures. However, the performance on FaceSwap (91.1%) is slightly lower compared to the other categories, a trend that aligns with findings from related methods such as LipForensics and RealForensics [29]). This could be attributed to FaceSwap's use of target face landmarks for generating source faces, which might result in more precise lip shapes. Nonetheless, our method demonstrates the significant capability in detecting such forgeries through contextual analysis, indicating its effectiveness against diverse manipulation techniques. Notably, our method significantly outperforms AVAD, which nearly produces random results. And we note that SpeechForensics-Local also achieves considerable performance on this dataset, although modeling local speech information.

Table 1: **Cross-manipulation generalization.** We report video-level AUC (%) on FF++, which contains four manipulation methods, i.e., Deepfakes (DF), FaceSwap (FS), Face2Face (F2F) and NeuralTextures (NT). * denotes results of our reproduction.

Method		Train on remaining three				
		DF	FS	F2F	NT	**Avg**
Supervised	Xception [55]	93.9	51.2	86.8	79.7	77.9
	Patch-based [12]	94.0	60.5	87.3	84.8	81.7
	Face X-ray [43]	99.5	93.2	94.5	92.5	94.9
	LipForensics [30]	99.7	90.1	99.7	99.1	97.1
	RealForensics* [29]	**100.**	96.1	99.5	97.0	98.1
	FTCN [71]	99.8	**99.6**	98.2	95.6	**98.3**
Unsupervised	AVAD* [24]	59.2	55.1	59.9	58.4	58.2
	SpeechForensics-Local	95.6	74.9	95.1	89.1	88.7
	SpeechForensics (ours)	99.4	91.1	**100.**	**100.**	97.6

Cross-Dataset Generalization. We further extend our evaluation to include a cross-dataset comparison, aligning with the practise in prior works [30, 29, 13]. This involves testing the performance of our method on the unseen datasets FakeAVCeleb [38] and KoDF[41], with the supervised models initially trained on the FF++ dataset. In addition, we also report the results of every category within the FakeAVCeleb dataset, which is segmented into five categories based on the manipulation techniques used Faceswap [40] (FS), FSGAN [50], Wav2Lip [53] (WL), Faceswap-Wav2Lip (FS-WL) and FSGAN-Wav2Lip (FSGAN-WL), with the latter two categories indicating the combined use of manipulation methods.

The results in table 2 show that our method significantly outperforms both supervised and unsupervised counterparts in the cross-dataset setting, outperforming previous state-of-the-art method, RealForensics, by 8.8% on FakeAVCeleb and 7.4% on KoDF. It is worth noting that SpeechForensics-Local fails to detect fake videos generated by Wav2Lip, as we mentioned above, suggesting the key role of global temporal information for forgery video detection. Conversely, AVAD [24] shows promise in detecting forgeries generated by Wav2Lip [53], though its performance on other types of forgery is yet to be fully assessed. Our approach, in contrast, delivers exceptional performance across all forgery types, achieving a perfect 100% on Wav2Lip-manipulated video. This underscores the fact that accurate lip synchronization at a local level does not necessarily imply global semantic integrity. We also provide more multimodal experiment results in appendix A.3.

Cross-Language Generalization. Considering the linguistic diversity encountered in real-world video content, we expand our evaluation to assess the cross-language generalization capabilities of

Table 2: **Cross-dataset generalization.** Video-level AUC (%) on FakeAVCeleb and KoDF. We report the results of every categories of FakeAVCeleb, and the overall performance on it is reported in **Overall**. The average performance over two datasets is reported in **Avg**.

	Method	FakeAVCeleb						KoDF	**Avg**
		FS	FSGAN	WL	FS-WL	FSGAN-WL	**Overall**		
Supervised	Xception [55]	67.0	62.5	59.7	57.2	68.0	61.6	77.7	69.7
	Patch-based [12]	97.4	80.5	78.9	93.8	87.8	83.6	83.9	83.8
	Face X-ray [43]	89.9	85.4	69.5	84.4	87.6	78.4	83.0	80.7
	LipForensics [30]	89.5	96.4	85.6	87.2	95.8	89.8	59.6	74.7
	FTCN[71]	89.3	79.9	80.6	85.2	86.1	82.3	76.5	79.4
	RealForensics [29]	**98.1**	**100.**	81.0	94.7	99.2	90.2	84.3	87.3
Unsupervised	AVAD [24]	52.8	53.9	93.9	95.8	94.3	85.0	58.0	71.5
	SpeechForensics-Local	69.3	85.4	0.10	0.08	0.08	19.0	48.3	33.7
	SpeechForensics (ours)	93.9	96.0	**100.**	**99.9**	**99.9**	**99.0**	**91.7**	**95.4**

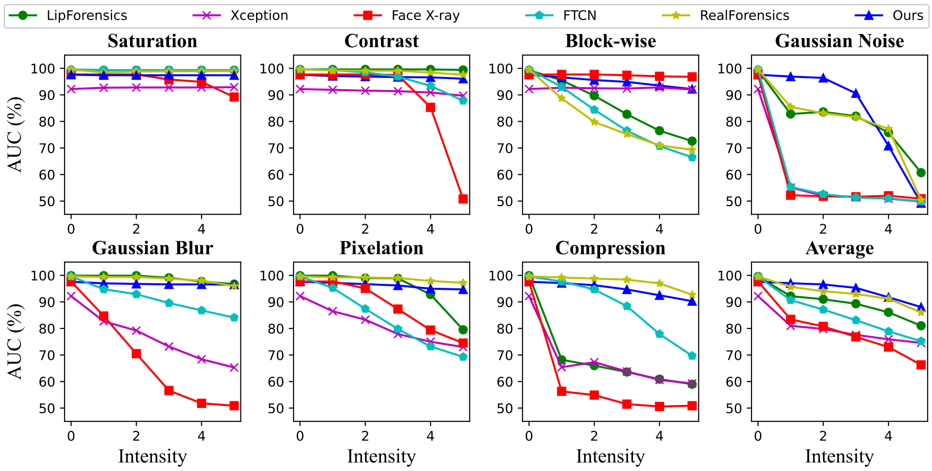


Figure 2: **Robustness to unseen perturbations.** Video-level AUC scores (%) are reported under different perturbations. Each perturbation contains five intensity levels [34]. "Average" denotes the mean of each perturbation under each intensity level.

our approach. To categorize the languages present in the FF++ dataset, we utilize `Whisper` [54] for language detection and subsequently split the videos into various language categories: including *English* (EN), *Arabic* (AR), *Spanish* (ES), *Russian* (RU), *Ukrainian* (UK), *Tagalog* (TL) and others. The results, presented in table 3, illustrate the AUC scores achieved for each language category. Our findings indicate that our method maintains effective performance across different languages, even though it was originally trained on datasets predominantly in *English*. This outcome underscores the versatility of the audio-visual speech representations learned by our model, demonstrating their language-agnostic nature and highlighting the method's potential applicability in diverse linguistic contexts.

Table 3: **Cross-language generalization.** AUC (%) scores on videos of different languages in the FF++.

Language	EN	AR	ES	RU	UK	TL	Others
AUC	97.8	98.3	98.3	100.	99.3	99.7	97.2

Robustness to Unseen Perturbations. Considering the prevalence of image post-processing operations on social media platforms, such as compression, the robustness of detection systems emerges as a crucial challenge. In line with previous studies [30, 29, 64, 24], we evaluate the robustness of our method against various perturbations in the FF++ dataset, and these include Color saturation change, Color contrast change, Block-wise distortion, Gaussian noise, Gaussian blur, Pixelation and Video compression, each applied at 5 different intensity levels as in [34].

The results in fig. 2 compare the performance of our method against five supervised baselines under these conditions.It can be found that detectors primarily relying on low-level texture features, such as `Face X-ray` [43], `Xception` [55] and `FTCN` [71], are vulnerable to most of the common perturbations. Notably, while `Face X-ray` and `Xception` show diminished effectiveness against

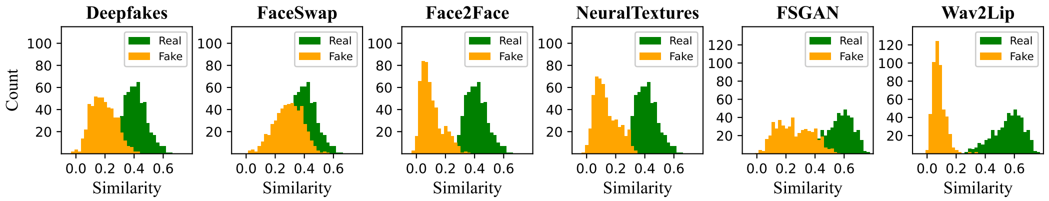
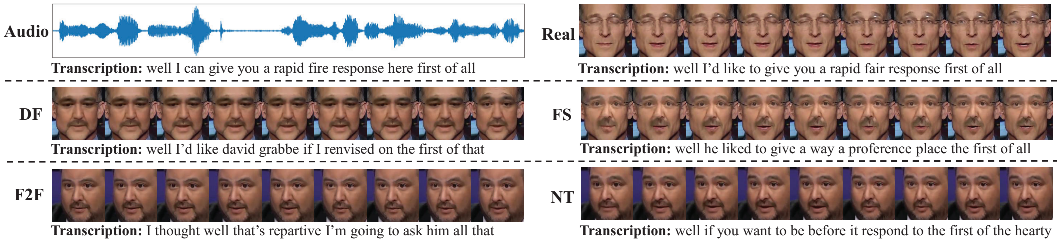


Figure 3: **Visualized analysis.** Cosine similarity distributions of audio and visual speech representations for real videos and fake videos generated by different manipulation methods.



Transcription: well I can give you a rapid fire response here first of all

Transcription: well I'd like to give you a rapid fair response first of all

Transcription: well I'd like david grabbe if I renvised on the first of that

Transcription: well he liked to give a way a proference place the first of all

Transcription: I thought well that's repartive I'm going to ask him all that

Transcription: well if you want to be before it respond to the first of the hearty

Figure 4: **Interpretative analysis.** The transcriptions are based on audio and visual speech representations of real and different types of fake videos. We show the transcriptions of each type of video containing the same audio.

perturbations that attenuate high-frequency content (e.g., *blur*, *compression*), FTCN is particularly sensitive to perturbations that disrupt temporal coherence (e.g., *noise*).

Besides, the three video-based methods, FTCN [71], LipForensics [30] and RealForensics [29], have all shown sensitivity to Block-wise distortion, suggesting they rely on low-level temporal features to some extent. Conversely, our method demonstrates exceptional robustness against all types of perturbations. Moreover, our approach consistently outperforms LipForensics, which also models lip movements, across various corruption scenarios, although starting from a lower point (97.6% vs 99.6%). This indicates that our method is capable of harnessing more potent semantic representations for the purpose of forgery detection, despite being trained exclusively on real data. We provide perturbation examples and more results in appendix A.4.

4.3 Qualitative Results

Visualization. To demonstrate the effectiveness of our method, we conduct an in-depth visual analysis utilizing FF++ and FakeAVCeleb datasets. Specifically, we included all forgery categories from FF++, i.e., Deepfakes, FaceSwap, Face2Face and NeuralTextures. From the FakeAVCeleb dataset, we focused on manipulations made with FSGAN and Wav2Lip, selecting a random sample of 500 videos. As a result, our experiment covers a total of six types of forgery.

For each type of forgery, we calculate the cosine similarity between visual and corresponding audio speech representations extracted from videos. Figure 3 shows the cosine similarity distribution for each category. As we can see, almost all types of fake videos are clearly differentiated from real videos around a cosine similarity threshold of 0.3. An interesting finding is that our method exhibits exceptional performance on face reenactment techniques, i.e., Face2Face, NeuralTextures and Wav2Lip, likely due to these methods prioritizing overall visual fidelity at the expense of accurate lip movements. More visualized results are in appendix A.5.

Interpretative Analysis. Exploring another intriguing aspect, we delve into understanding how our method functions and what the derived audio-visual representations signify. To shed light on this, we conduct an interpretative analysis on the FF++. Specifically, we transcribe the extracted representations using an audio-visual speech recognition model as proposed by [56], fine-tuned for lipreading with the pretrained audio-visual speech representation model. This enabled us to transcribe specific speech content from both the lip movements and corresponding audio segments in real and fake videos, as shown in fig. 4. Note only the mouth region sequences are fed into the model to get the transcriptions. As can be seen, the sentence transcribed from real lip movement frames is close to the sentence transcribed from audio. In contrast, lip movements in fake videos frequently result in

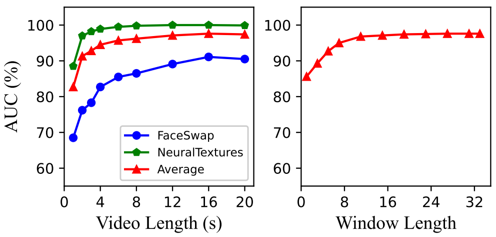


Figure 5: **Influence of video length and sliding-window length.** We evaluate the performance of our method conditioned on different input lengths and sliding-window lengths.

Table 4: **Effect of different models and time offset assumptions.** We report the performance of models with different architectures and training datasets on FF++ and FakeAVCeleb.

Model	Offset	Backbone	Dataset	FF++	FakeAVCeleb
AVHuBERT [56]	Fixed	BASE	LRS3	95.3	97.0
		BASE	LRS3+Vox2	96.1	97.9
		LARGE	LRS3	95.7	96.8
		LARGE	LRS3+Vox2	**97.6**	99.0
	Dynamic	LARGE	LRS3+Vox2	93.7	98.7
VATLM [73]	Fixed	LARGE	LRS3+Vox2	97.1	**99.3**

nonsensical or chaotic transcriptions. It suggests that these audio and visual speech representations indeed contain semantic information, which can be used for forgery detection.

4.4 Ablation study

Influence of Video Clip Length. Given that our model accommodates video clips of varying lengths, we investigated how this aspect impacts the performance of our method on the FF++ dataset. For this purpose, we selected `FaceSwap` and `Face2Face` as two emblematic types of forgeries and segmented videos into various durations, ranging from 1 second to 20 seconds, while maintaining consistency in all other hyperparameters. Results in fig. 5 indicate a clear trend: the performance of our method is continuously enhanced with the extension of the video length to 16 seconds, suggesting long-range temporal inconsistencies exist in forgery videos. While previous detectors, e.g.,FTCN [71] and RealForensics [29], can only utilize short-range temporal features, resulting in their suboptimal performance.

Different time offset assumptions. We also study the effect of different time offset assumptions between audio and visual streams, i.e., fixed and dynamic. As shown in table 4, the dynamic based method achieves slightly better performance based on the fixed time offset assumption. A reasonable explanation is that, compared to real videos, the time offsets of fake videos will be more inconsistent due to the uncertainty of the forgery process. And the DTW algorithm makes fake videos have higher matching scores, which is unfavorable to the forgery detection. Furthermore, we investigate the influence of different maximum offset τ, corresponding sliding-window length $2\tau + 1$. As fig. 5 shows, increasing window length brings improved performance, reaching a maximum of around 31.

Different Models and Datasets. We further evaluate the effect of different models and training datasets. For AVHuBERT [56], we use models with two configurations: BASE, which comprises 12 transformer blocks, and LARGE, which includes 24 transformer blocks. Each configuration was trained on two datasets: `LRS3` alone and a combination of `LRS3` and `VoxCeleb2`, respectively. Furthermore, we evaluate another model, VATLM [73], which also fits our framework but incorporates text modality. The experimental results on `FF++` and `FakeAVCeleb` are shown in table 4. Results show larger models and more training data both boost the performance of our approach. While compared with model size, the influence introduced by datasets is more pronounced. And both AVHuBERT and VATLM obtain remarkable results.

5 Conclusion and Discussion

In this paper, we have developed a method for forgery detection that identifies discrepancies between audio and visual speech representations. Demonstrating exceptional generalization capabilities to unseen manipulations and robustness against prevalent perturbations, our approach sets a new benchmark, notably without relying on fake videos for training. It also eliminates the need for finetuning and downstream tasks, significantly streamlining the detection workflow. Moreover, since our method is based on speech representation learning, it can be implemented in a training-free manner and may achieve better performance as the latter advances. We are optimistic that our contributions will inspire further advancements in the field of forgery detection research.

Limitations. While our method exhibits robust performance across diverse evaluations, it is not without its limitations. Primarily, it is constrained by its reliance on visual speech representations derived from lip movements, rendering it unsuitable for detecting forgeries that do not alter mouth regions. However, we note that facial forgeries typically involve the mouth area. In addition, it may suffer a certain level of performance degradation when encountering extreme testing samples, e.g., videos with numerous silent clips or audio signals containing significant amount of ambient noise, e.g., background music. These considerations highlight areas for potential improvement and future exploration in enhancing the versatility and applicability of forgery detection techniques.

Broader Impacts. Our work is aimed at fighting against face forgery technologies. And we hope it could encourage more future detection works. However, since forgery and detection are two game-playing technologies, the emergence of new detection methods may lead to the evolution of forgery methods. And we suggest that detection systems integrate different detection methods to combat potential new face forgery methods.

6 Acknowledgement

This work has been supported by the National Key R&D Program of China (Grant No. 2021YFF0602104).

References

[1] Deepfakes. `https://github.com/deepfakes/faceswap`. [Accessed: 2024-4-25].

[2] Deepfakesfaceswap. `https://github.com/deepfakes/faceswap`. [Accessed: 2024-4-25].

[3] Faceswap. `https://github.com/MarekKowalski/FaceSwap`. [Accessed: 2024-4-25].

[4] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018.

[5] Triantafyllos Afouras, Yuki M Asano, Francois Fagan, Andrea Vedaldi, and Florian Metze. Self-supervised object detection from audio-visual correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10575–10586, 2022.

[6] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727, 2018.

[7] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018.

[8] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[9] Weiming Bai, Yufan Liu, Zhipeng Zhang, Bing Li, and Weiming Hu. Aunet: Learning relations between action units for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24709–24719, 2023.

[10] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*, pages 1021–1030, 2017.

[11] Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, and Munawar Hayat. Marlin: Masked autoencoder for facial video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1493–1504, 2023.

[12] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 103–120. Springer, 2020.

[13] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18710–18719, 2022.

[14] Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Xiaojun Chang, and Liqiang Nie. Voice-face homogeneity tells deepfake. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–22, 2023.

[15] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. Not made for each other-audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the 28th ACM international conference on multimedia*, pages 439–447, 2020.

[16] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

[17] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 251–263. Springer, 2017.

[18] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Trans on Pattern Analysis and Machine Intelligence*, 2020.

[19] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018.

[20] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

[21] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pages 5781–5790, 2020.

[22] Oscar De Lima, Sean Franklin, Shreshtha Basu, Blake Karwoski, and Annet George. Deepfake detection using spatiotemporal convolutional networks. *arXiv preprint arXiv:2006.14749*, 2020.

[23] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020.

[24] Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10491–10503, 2023.

[25] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, and Lizhuang Ma. Delving into the local: Dynamic inconsistency learning for deepfake video detection. In *AAAI*, volume 36, pages 744–752, 2022.

[26] Zhihao Gu, Taiping Yao, C Yang, Ran Yi, Shouhong Ding, and Lizhuang Ma. Region-aware temporal inconsistency learning for deepfake video detection. In *IJCAI*, volume 1, 2022.

[27] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[28] Alexandros Haliassos, Pingchuan Ma, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Jointly learning visual and auditory speech representations from raw data. In *International Conference on Learning Representations*, 2023.

[29] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja 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*, pages 14950–14962, 2022.

[30] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5039–5049, 2021.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[32] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[33] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in Neural Information Processing Systems*, 31, 2018.

[34] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2889–2898, 2020.

[35] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[36] Mikolaj Kegler, Pierre Beckmann, and Milos Cernak. Deep speech inpainting of time-frequency masks. 2020.

[37] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.

[38] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[39] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[40] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3677–3685, 2017.

[41] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10744–10753, 2021.

[42] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6458–6467, 2021.

[43] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020.

[44] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018.

[45] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216, 2020.

[46] Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. Auto-avsr: Audio-visual speech recognition with automatic labels. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[47] Pingchuan Ma, Rodrigo Mira, Stavros Petridis, Björn W Schuller, and Maja Pantic. Lira: Learning visual speech representations from audio through self-supervision. *arXiv preprint arXiv:2106.09171*, 2021.

[48] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7613–7617. IEEE, 2021.

[49] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[50] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7184–7193, 2019.

[51] Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. Avff: Audio-visual feature fusion for video deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27102–27112, 2024.

[52] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, et al. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*, 2020.

[53] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492, 2020.

[54] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

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

[56] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *International Conference on Learning Representations*, 2022.

[57] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022.

[58] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *NIPS*, 32, 2019.

[59] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[60] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016.

[61] Suramya Tomar. Converting video formats with ffmpeg. *Linux journal*, 2006(146):10, 2006.

[62] Chengrui Wang and Weihong Deng. Representative forgery mining for fake face detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14923–14932, 2021.

[63] Xueyu Wang, Jiajun Huang, Siqi Ma, Surya Nepal, and Chang Xu. Deepfake disrupter: The detector of deepfake is my friend. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14920–14929, 2022.

[64] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4129–4138, 2023.

[65] Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren. Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security*, 18:2015–2029, 2023.

[66] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*, 2020.

[67] Zhaoyang Zeng, Daniel McDuff, Yale Song, et al. Contrastive learning of global and local video representations. *Advances in Neural Information Processing Systems*, 34:7025–7040, 2021.

[68] Daichi Zhang, Chenyu Li, Fanzhao Lin, Dan Zeng, and Shiming Ge. Detecting deepfake videos with temporal dropout 3dcnn. IJCAI, 2021.

[69] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Weiming Zhang, and Nenghai Yu. Self-supervised transformer for deepfake detection. *arXiv preprint arXiv:2203.01265*, 2022.

[70] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15023–15033, 2021.

[71] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15044–15054, 2021.

[72] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14800–14809, 2021.

[73] Qiushi Zhu, Long Zhou, Ziqiang Zhang, Shujie Liu, Binxing Jiao, Jie Zhang, Lirong Dai, Daxin Jiang, Jinyu Li, and Furu Wei. Vatlm: Visual-audio-text pre-training with unified masked prediction for speech representation learning. *IEEE Transactions on Multimedia*, 2023.

A Appendix

A.1 Architecture Details.

The visual frontend is a modified ResNet-18 [48], where the first convolutional layer is substituted by a 3D convolutional layer. The resulting visual features are then transformed into 1024-dimensional tensors for each input frame through the application of 2D global average pooling. See table 5 for more details. The masked predictor consists of 24 standard transformer encoder blocks, each featuring 16 attention heads and 1024 channels. A final linear projection layer, with an output dimension of 256, is employed to deduce the ultimate cluster assignments.

Table 5: **Visual frontend architecture.**The output size is of the form $T \times H \times W$, where T denotes the number of input frames, H denotes the height of frames and W denotes the width.

stage	filters	output size
conv_1	$5 \times 7 \times 7$, stride $1 \times 2 \times 2$	$T \times 44 \times 44$
pool_1	max, $1 \times 3 \times 3$, stride $1 \times 2 \times 2$	$T \times 22 \times 22$
res_1	$\begin{bmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{bmatrix} \times 2$	$T \times 22 \times 22$
res_2	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 2$	$T \times 11 \times 11$
res_3	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	$T \times 6 \times 6$
res_4	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 2$	$T \times 3 \times 3$
pool_2	global spatial average pool	$T \times 1 \times 1$

A.2 Compared Baselines

We compare our method with both supervised and unsupervised methods.

Supervised Methods. The state-of-the-art supervised methods include: 1) **Xception** [55]: a widely used baseline for generalization comparison. 2) **Patch-based** [12]: a patch-based forgery detection model with local receptive fields. 3) **Face X-ray** [43]: a detector focusing on blending boundaries in fake images. 4) **LipForensics** [30]: it targets unnatural mouth movements existing in forgery videos. 5) **FTCN** [71]: a video detector modeling temporal features via special architecture. 6) **RealForensics** [29]: it aims to learn temporally dense representations of facial movements via audio-visual self-supervision, which facilitates the generalization of forgery detectors.

Unsupervised Methods. 1)**AVAD** [13]: it leverages a pre-trained audio-visual synchronization network as its core framework, and adopts a downstream autoregressive model to learn the distribution of time delays between visual and auditory signals. 2)**SpeechForensics-Local**: We utilize two Resnet 2D [31] models to extract audio and visual speech representations, respectively. The visual input contains 5 successive frames and the audio encoder extract features from 0.2s audio clips. We also train it on the `LRS3` [7] dataset, and we dub it as `SpeechForensics-Local`, since it learns local phoneme information.

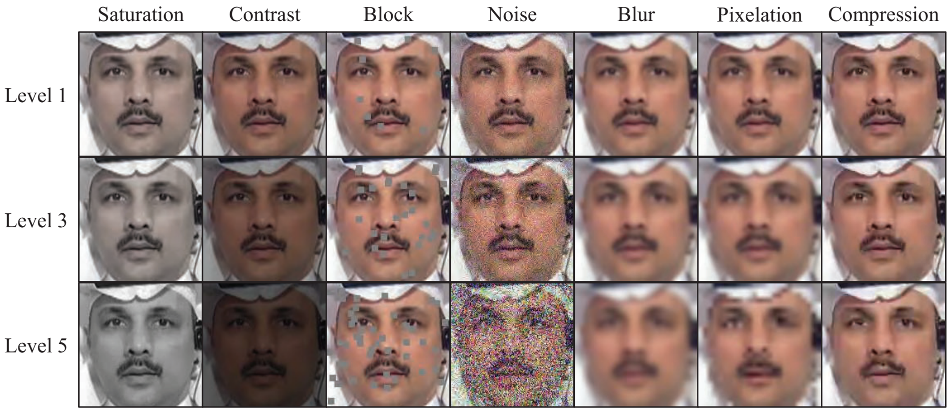


Figure 6: **Perturbed examples.** Visualization of all types of perturbations at different intensity levels. We present three representative (mild, moderate and severe) intensity levels.

Table 7: **Average robustness to unseen perturbations.** The average AUC (%) scores of different methods over each type of perturbations at all intensity levels. **Avg:** average performance on corrupted videos. **Drop:** decreased performance compared to clean videos.

Method	Clean	Saturation	Contrast	Block	Noise	Blur	Pixelation	Compress	**Avg/Drop**
Xception [55]	92.2	92.8	91.1	92.5	51.9	73.8	79.1	63.3	77.8/-14.4
Face X-ray [43]	97.7	95.1	85.8	**97.3**	51.7	62.9	86.8	52.8	76.1/-21.6
LipForensics [30]	99.6	99.2	**99.6**	83.4	77.0	**98.7**	94.0	63.5	87.9/-11.7
FTCN [71]	99.4	**99.3**	95.2	78.2	52.0	89.6	81.0	85.7	83.0/-16.4
RealForensics[29]	99.6	98.7	98.7	76.8	75.6	98.3	**98.5**	**97.2**	92.0/-7.6
Ours	97.6	97.4	96.7	94.6	**80.8**	97.4	96.7	94.2	**94.1/-3.5**

A.3 More Comparisons with Multimodal Baselines

We also compare our method with more multimodal, i.e., audio-visual, baselines under the cross-dataset setting. Since they do not have open source codes or pre-trained weights, we provide their results on the FakeAVCeleb according to [65].

Table 6: **Cross-dataset generalization comparisons with multimodal baselines.** We report the AUC (%) scores on the FakeAVCeleb.

Method	MDS [15]	VFD [14]	Avoid-DF [65]	AVAD [24]	Ours
	Supervised	Supervised	Supervised	Unsupervised	Unsupervised
AUC	76.7	82.5	85.8	85.0	**99.0**

As shown in table 6, our method significantly outperform both supervised and unsupervised counterparts. We note that many audio-visual face forgery detection methods adopt the cross-modal fusion strategy [72, 65, 51], However, our method prove the cross-modal fusion may not be necessary for the face forgery detection task, which we believe needs to be explored further.

A.4 Robustness Experiments

Following [30], we apply the perturbations using the DeeperForensics [34] code[3], which implements seven different perturbations conditioned on five intensity levels. And we present some perturbation examples in fig. 6.

In table 7, we report the average AUC scores across all intensity levels for each perturbation. It can be seen our method achieves the state-of-the-art robustness to unseen perturbations, although it starts from a lower AUC score than other supervised methods, e.g., LipForensics [30] and FTCN [71].

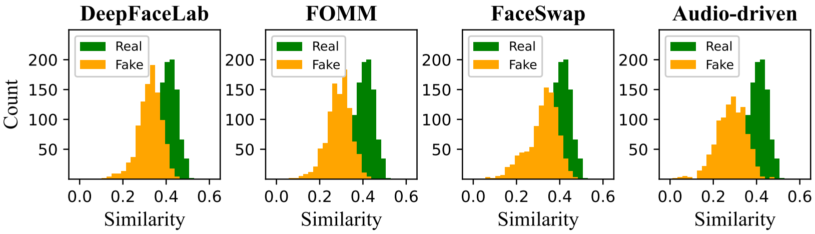


Figure 7: **Visualization of cosine similarity distribution of KoDF.** Cosine similarity distribution of audio and visual speech representations of different types of videos in KoDF.

A.5 Visualization Analysis

In fig. 7, we show the cosine similarity distribution of different types of forgeries of KoDF [41], involving DeepFaceLab [52], FOMM [58], FaceSwap [2] and Audio-driven (including ATFHP [66] and Wav2Lip [53]). It shows our method has strong generalization to unseen datasets and languages.

[3] https://github.com/EndlessSora/DeeperForensics-1.0/tree/master/perturbation

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have faithfully described the motivations and performance of our method.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our method in the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed information about the models and datasets in the experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have released the source code and execution steps in the supplementary material and an anonymous repository. And we will open the repository once our paper gets published.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify the key configurations about the training and test details in the experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our experiments use a empirical random seed and does not report error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Current focus of the research domain of this work is on accuracy rather than speed. Thus we mainly report the experimental results about accuracy.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform with NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We point out the potential societal impacts of this work in the conclusion section.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The FF++ dataset is released under the FaceForensics Terms of Use. The FakeAVCeleb dataset is released under the Terms of Use FakeAVCeleb. The KoDF dataset is released under the KoDF Terms of Use. The AVHuBERT models are released under Terms of Use AVHuBERT.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper dose not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.