Unsupervised Multimodal Deepfake Detection Through Explicit Intra-Modal and Cross-Modal Inconsistency Discovery

Mulin Tian[1,2,*]
mulintia@usc.edu
Mahyar Khayatkhoei[1,*]
mkhayat@isi.edu
Joe Mathai[1,*]
jmathai@isi.edu
Wael AbdAlmageed[3]
cwabdalm@clemson.edu

[1] Information Sciences Institute
University of Southern California
Marina del Rey, USA

[2] Hsieh Department of Electrical and Computer Engineering
University of Southern California
Los Angeles, USA

[3] Holcombe Department of Electrical and Computer Engineering
Clemson University
Clemson, USA

Abstract

Deepfake videos present an increasing threat to society with potentially negative impact on criminal justice, democracy, and personal safety and privacy. Meanwhile, detecting deepfakes, at scale, remains a very challenging task that often requires labeled training data from existing deepfake generation methods. Further, even the most accurate supervised deepfake detection methods do not generalize to deepfakes generated using new generation methods. In this paper, we propose a novel unsupervised method for detecting deepfake videos by explicitly discovering intra-modal and cross-modal inconsistencies between video segments. The fundamental hypothesis behind our proposed method is that motion or identity inconsistencies are inevitable in motion-transfer deepfake videos. We provide the first mathematical and empirical evidence to support this hypothesis, hence justifying the generalizability of our method. Our proposed method outperforms prior state-of-the-art unsupervised deepfake detection methods on the challenging FakeAVCeleb dataset, and also has several additional advantages: it is scalable because it does not require pristine (real) samples for each identity during inference and therefore can apply to arbitrarily many identities; generalizable because it is trained only on real videos and therefore does not rely on any particular deepfake method; reliable because it does not rely on any likelihood estimation in high dimensions; and explainable because it can pinpoint the exact location of modality inconsistencies.

*These authors contributed equally to this work.

1 Introduction

The rapid advancement of generative deep learning [26, 43, 44], fueled by faster and cheaper compute power and exploding data availability, has blurred the line between fact and fiction. In particular, *deepfakes*[*]—videos in which the motion of a source video is transferred to a target identity such that it appears to say the words uttered by the source—are becoming increasingly hard to distinguish from real videos [15]. Since existing deepfake detection methods themselves can be used as part of the objective in the generation process to improve the deepfake quality, this rapid advancement in generative deep learning leads to a daunting question: *will machines eventually be able to imitate any person without leaving any trace?* Despite the alarming empirical evidence [3, 24], we conjecture that the answer could be negative; that is, deepfakes might always contain a *detectable inevitable trace*. The intuition behind this conjecture is illustrated in fig. 1. While the synthesized video (middle row) appears very realistic in each frame and its motion almost exactly matches those of the source, a closer look reveals that at certain frames the identity of the person has noticeably changed (depicted by a red dashed box). We conjecture that this phenomenon, which occurs in various deepfake videos [27], is a consequence of a fundamental property of facial motion and identity, that *facial motion and identity are not independent variables*. Consequently, a deepfake model must either not perfectly transfer the motion (leading to motion inconsistencies in the target video) or perfectly transfer the motion and partially transfer the identity as well (leading to identity inconsistencies in the target video). **In this paper, we develop a new method for explicitly detecting these conjectured inconsistencies (section 3), and provide mathematical and empirical evidence to support our conjecture (in the supplementary material C.1).**

Figure 1: When transferring the motion of the source video *(top)* to the target identity *(Angelina Jolie)*, the deepfake generation method [15] faces a trade-off: *(middle)* matching motion exactly results in frames with wrong identity which can be detected by looking for intra-modal identity inconsistency; or *(bottom)* matching identity exactly results in motion distortion which can be detected by looking for video cross-modal inconsistency with audio (*e.g.*, the lips do not move at moments where audio magnitude shows speaking, and vice versa). Red boxes indicate inconsistent segments.

Our proposed *unsupervised* deepfake detection method is designed to *explicitly* detect intra- and cross-modal inconsistencies within a given video. While extracting consistency-based *features* have been explored in prior deepfake detection (*i.e.*, learning joint features from video-audio based on a contrastive loss [4, 8]), *our work is the first to explicitly pin-*

[*]We use *deepfake* as an umbrella term to refer to fake video generation in general including face-swapping, face reenactment, face generation, etc.

point which parts of a given video are inconsistent, and it achieves this without relying on direct comparison to any pristine (real) video. Our method outperforms prior unsupervised detection methods on the challenging FakeAVCeleb [27] dataset (achieving a new best average AUC of 96.81%). We summarize the novelty and benefits of our method compared to prior works below.

Generalizability: Compared to supervised methods [9, 13, 32, 36], which rely on existing deepfake generation models, our method is unsupervised and is trained solely on real videos, thereby does not rely on the particular artifacts of existing deepfakes.

Scalability: Compared to POI-Forensics [8], which requires access to a set of pristine reference samples for each identity to compare against during inference, our method does not require any pristine samples during inference, therefore can better scale to many new identities.

Reliability: Compared to AV-Anomaly [17], which relies on generative modeling to estimate the likelihood of real data (making it susceptible to the known unreliability of likelihood estimation in high-dimensions [38]), our method does not use any likelihood estimation, and instead directly compares a given video with itself to find mismatching segments within and across modalities.

Explainability: Our method is explainable by design, because in order to detect whether a given video is fake, it must discover that a portion of the video is inconsistent with another portion of the same video. These two inconsistent portions can be provided to a human expert as an explanation for verification.

2 Related Works

Visual Deepfake Detection. The visual-only deepfake detection methods focus on detecting the artifacts in the visual part of a video. Li *et al*. [33] argued that the generated face is often warped to match the original faces in the source video, and proposed a method to detect artifacts caused by this face warping on images using a convolution neural network (CNN). Güera *et al*. [20] argued for the importance of temporal clues for deepfake detection, and proposed a temporal-aware supervised method which trains a recurrent neural network (RNN) on per-frame features extracted by a CNN to detect deepfake videos. Yang *et al*. [48] argued that generated images often have incorrect 3D head poses, and proposed a supervised method for detecting deepfakes based on 3D pose estimation.

Multimodal Deepfake Detection. Several approaches have recently focused on multimodal deepfake detection by utilizing both visual and auditory information. Li *et al*. [37] proposed detecting deepfake videos by analyzing audio-visual cues and perceived emotions which was extracted from Memory Fusion Network (MFN) [50]. AV-Anomaly [17] first learns a joint audio-visual segment feature in a contrastive setting from real videos, and then trains a likelihood estimator on those features extracted from real videos (an autoregressive generative model of sequences of synchronization characteristics). The likelihood estimator is then used to score a query video's likelihood of being real. POI-Forensics [8] proposed an audio-visual identification verification approach, where a model is trained to extract features that indicate the identities of videos through contrastive learning on real videos, and then during inference, these features are used to compare a query video to a set of existing pristine (real) videos, to determine a fakeness score. AVFF [40] recently proposed a two-stage model that first trains a joint audio-visual representation for real videos through unsupervised learning, and then trains a classifier on this representation for deepfake detection.

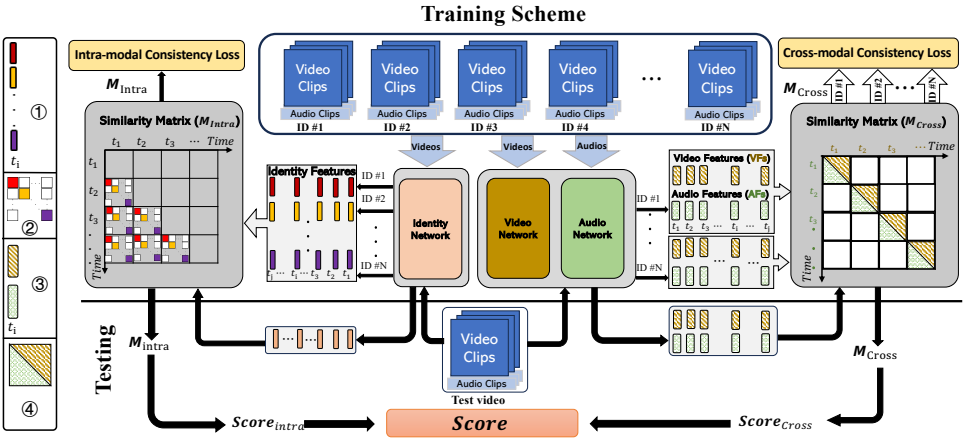

Figure 2: Training and testing scheme for intra-modal consistency and cross-modal consistency methods. For each training batch, we take multiple fixed-size video and audio clips of N distinct identities and feed them into our networks. ① is an output (feature vector) of all identities extracted from the identity network at time-window t_a, The similarity matrix computed on time dimension is given in the gray box on the left, each element ② represents the similarity matrix of ① on a specific time-window pair (t_a, t_b). A intra-modal consistency loss for identity network training is calculated based on this tensor. ③ denotes the feature vector generated by video and audio network at time-window t_a. The features of each individual across multiple time windows are used to generate their corresponding similarity matrix. A cross-modal consistency loss for video and audio network training is calculated from these N 2-dimensional matrices.

In contrast to the aforementioned methods which use audio-visual consistency objectives as a way to learn representative features for later processing (likelihood estimation or comparison to pristine videos or supervised classifier training), we aim to develop a method that is purely based on unsupervised learning and can explicitly identify inconsistencies in a query video. We also provide theoretical justification for the generalization of the proposed method to unseen deepfake methods.

3 Deepfake Detection

Motivated by the issue illustrated in fig. 1, we propose a deepfake detection method based on the premise of inevitable inconsistencies in terms of (1) identity inconsistency in the generated video by comparing different video segment pairs in the video, and (2) motion distortions in the generated video by comparing video and audio segments. Our method is summarized in fig. 2, and described in detail in this section.

Notation and Training Setup. Assume a training batch that includes multiple videos $\{v_i\}_{i=1}^{N}$ for N distinct *real* identities, each video having T different time-windows (a time-window is a fixed duration of time). Each video goes through three mappings parameterized by neural networks to extract identity, visual and audio features. Let $\mu : \mathcal{V} \times \{1 \dots T\} \rightarrow \{y \in \mathbb{R}^d : ||y||_2 = 1\}$ define a mapping between a video and its corresponding identity fea-

tures at a time-window $t \in [1,T]$, similarly $\gamma : \mathcal{V} \times \{1 \dots T\} \rightarrow \{y \in \mathbb{R}^d : ||y||_2 = 1\}$ define a mapping between a video and its corresponding visual features at a time-window, and $\alpha : \mathcal{V} \times \{1 \dots T\} \rightarrow \{y \in \mathbb{R}^d : ||y||_2 = 1\}$ define a mapping between a video and its corresponding audio features at a time-window. Whenever referring to a feature extracted from the video of identity i at time-window t, we summarize $\mu(v_i,t)$ into $\mu_i(t)$ for brevity. The dot product is denoted by $\langle \cdot, \cdot \rangle$.

Intra-modal Consistency Loss. In order to detect identity inconsistencies, we must learn an identity feature extractor μ that is sensitive to slight changes of identity within any video. Therefore, we propose the intra-modal consistency loss, whose objective is to guide μ to learn maximally divergent feature vectors between different identities, and maximally convergent feature vectors between different observations/samples of the same identity. We use μ to extract identify features for each sample in the batch as illustrated in the left side of fig. 2. We then measure the similarity between all pairs of identity vectors $\langle \mu_i(t), \mu_j(q) \rangle$, where i, j are identity indices and t, q time-window indices, resulting in a $T \times T \times N \times N$ similarity tensor. eq. (1) shows the overall *intra-modal consistency loss* $L_{\odot}$:

$$L_{\odot} = \frac{-1}{N \times T^2} \sum_{t=1}^{T} \sum_{q=1}^{T} \sum_{i=1}^{N} \log \left(\frac{\exp(\langle \mu_i(t), \mu_i(q) \rangle / \tau)}{\sum_{j=1}^{N} \exp(\langle \mu_i(t), \mu_j(q) \rangle / \tau)} \right) \tag{1}$$

where τ is the temperature used to control the scale of similarity measurements. The left part of fig. 2 illustrates how we find the dissimilarity with given features from different identities (small white squares in ②-fig. 2) and the similarity with given features from the same identity (small colored squares in ②-fig. 2). The objective is to make the colored dot-products increase and white dot-products decrease such that vectors of similar identities becoming closer to each other than to vectors of dissimilar identities, across all pairs of time-windows. At inference time, the video is divided into T fixed sized windows. As illustrated by the black arrows in the left part of the testing flow in fig. 2, we use the trained μ to compute the $T \times T$ similarity matrix of identity features. The intra-modal consistency score of the test video using intra-modal consistency method ($Score_{\odot}$) is the n^{th} percentile of this similarity matrix.

Cross-modal Consistency Loss. In order to detect motion inconsistencies, we propose to learn a visual feature extractor γ and an audio feature extractor α that are sensitive to slight mismatches between audio and video of a speaking person. To that end, we construct the cross-modal consistency loss, whose objective is to guide γ and α to learn maximally divergent feature vectors between video and audio at different time-windows, and maximally convergent feature vectors between video and audio at the same time-windows. We use γ and α to extract visual and audio features for each sample in the batch as illustrated in the right side of fig. 2. We then measure the similarity between all pairs of audio-video feature vectors $\langle \gamma_i(t), \alpha_i(q) \rangle$, where i is the identity index and t, q are time-window indices, resulting in a $N \times T \times T$ similarity tensor. eq. (2) shows the overall *cross-modal consistency loss* $L_{\otimes}$:

$$L_{\otimes} = \frac{-1}{N \times T} \sum_{i=1}^{N} \sum_{t=1}^{T} \log \left(\frac{\exp(\langle \gamma_i(t), \alpha_i(t) \rangle / \tau')}{\sum_{q=1}^{T} \exp(\langle \gamma_i(t), \alpha_i(q) \rangle / \tau')} \right) + \log \left(\frac{\exp(\langle \gamma_i(t), \alpha_i(t) \rangle / \tau')}{\sum_{q=1}^{T} \exp(\langle \gamma_i(q), \alpha_i(t) \rangle / \tau')} \right) \tag{2}$$

where τ' is the temperature used to control the scale of similarity measurements. The right part of fig. 2 intuitively illustrates how we find the dissimilarity with given audio and video features at different time-windows (white squares in the gray block-fig. 2) and the similarity with given audio and video features from the same time-window (squares with green and

Table 1: AUC (%) and AP (%) of our methods (Intra-modal, Cross-modal, and Intra-Cross-modal) on FakeAVCeleb[27]. Best and second-best results of supervised and unsupervised methods are in bold and underlined. Our combined method outperforms the unsupervised state-of-the-art, and reaches within 2.7 absolute percentage points of best supervised method on average (AVG-FV). It also exceeds the unsupervised state-of-the-art method by 22.4 and the supervised state-of-the-art method by 7 percentage points in the RVFA category.

	Method	Modality	Pretrained Dataset	Catogory									
				RVFA		FVRA-WL		FVFA-FS		FVFA-GAN		FVFA-WL		AVG-FV	
				AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC
Supervised	Xception [4]	V	ImageNet	-	-	**97.8**	97.7	99.9	99.9	61.5	68.1	98.6	98.7	89.4	91.1
	LipForensics [4]	V	LRW	-	-	97.0	97.4	99.6	99.7	58.4	55.4	**100.**	**100.**	88.8	88.1
	AD DFD [4]	AV	Kinetics	74.9	73.3	96.2	97.4	**100.**	**100.**	77.4	78.3	95.6	96.5	92.3	93.1
	FTCN [4]	V	-	-	-	88.8	93.0	99.3	99.1	99.8	99.8	93.4	96.7	95.3	97.1
	RealForensics [4]	V	LRW	-	-	**93.3**	**92.4**	94.8	**98.2**	**100.**	**100.**	**99.9**	**100.**	99.4	99.8
	AVFF [4]	AV	LRS3 & FakeAvCeleb	**93.3**	**92.4**	94.8	**98.2**	**100.**	**100.**	**99.9**	**100.**	99.4	99.8	**98.5**	**99.5**
Unsupervised	AVBYOL [4]	AV	LRW	50.0	50.0	73.4	61.3	88.7	80.8	60.2	33.8	73.2	61.0	73.9	59.2
	VQ-GAN [4]	V	LRS2	-	-	50.3	49.3	57.5	53.0	49.6	48.0	62.4	56.9	55.0	51.8
	A-V Anomaly [4]	AV	LRS2	62.4	71.6	93.6	93.7	95.3	95.8	94.1	94.3	93.8	94.1	94.2	94.5
	A-V Anomaly [4]	AV	LRS3	70.7	80.5	91.1	93.0	91.0	92.3	91.6	92.7	91.4	93.1	91.3	92.8
	Ours (Intra-modal)	V	VoxCeleb2	-	-	65.14	67.99	86.11	86.65	90.08	90.65	63.28	66.15	76.15	77.86
	Ours (Cross-modal)	AV	VoxCeleb2	**99.68**	**99.65**	94.56	95.89	93.52	95.66	92.16	94.58	**95.29**	**96.25**	93.88	95.62
	Ours (Intra-Cross-modal)	AV+V	VoxCeleb2	98.49	99.41	**94.92**	**95.96**	**96.90**	**97.71**	**96.67**	**97.59**	94.84	95.99	**95.74**	**96.81**

brown filling as marked as ④-fig. 2). The objective is to make the colored dot-products increase and white dot-products decrease such that vectors of corresponding audio-visual time-windows become closer to each other than to all other audio or video vectors of different time-window for each identity. At inference/testing time, the testing video is divided into T fixed sized windows, and following the black arrows in the right part of the testing flow in fig. 2, we will use the trained γ and α to compute the $T \times T$ similarity matrix of visual-audio features. The score of the testing video ($Score_{\otimes}$) is determined by calculating the average value of its diagonal elements. We take the sum of our two scores $Score = Score_{\otimes} + Score_{\odot}$ as the final deepfake detection score. We visualize the inference of our methods on a real and a fake video in fig. 3 in supplementary material, and illustrate its inherent explainability.

4 Experimental Evaluations

Training Datasets. We use VoxCeleb2 [7] for training our methods, because our training process requires a substantial amount of speaking videos from various individuals, and VoxCeleb2 meets the requirements. VoxCeleb2 includes over 6000 different identities, each with videos of that identity speaking in multiple scenarios. Leveraging this characteristic, it is feasible to construct a substantial number of video and audio windows with fixed size. For ablation purposes, we also report the results of our method when trained on LRS3 [1] and without any training in the supplementary material.

Evaluation Datasets. We use FakeAVCeleb [27] – the most common evaluation benchmark used by recent multimodal deepfake detection methods [8, 17, 40] – to evaluate the performance of our model, which consists of 500 identities taken from VoxCeleb2, where both the visual and audio content of the videos of these identities is tampered. We excluded all videos of the 500 identities present in FakeAVCeleb from our training dataset (VoxCeleb2) in order to prevent our method from learning features specific to these identities. FakeAVCeleb dataset allows us to thoroughly test our model's ability to discern between different modalities of forgery. Specifically, FakeAVCeleb utilizes Wav2Lip [41], faceswap

[27] and fsgan [39] for video manipulation, along with SV2TTS [25] for audio manipulation. The manipulated videos are categorized into four groups: RealVideo-RealAudio, RealVideo-FakeAudio, FakeVideo-RealAudio, and FakeVideo-FakeAudio. The first two categories contain 500 videos each, and the others have about 10000 videos each. In addition, we also use KoDF [29], a deepfake dataset of Korean speaking videos, for evaluating the generalization of our method to people speaking in other languages (which indicates out-of-domain generalization to audio features).

Models. We use the AdaFace [28] model pre-trained on MS1MV2 [11], MS1MV3 [12] and WebFace4M [53] to extract per-frame identity features and use a transformer encoder to aggregate the identity in a time-window for our visual features (both μ and γ). For the audio features (α) we use the Whisper [42] encoder pre-trained on 680k hours of labeled audio-text data as our audio feature extractor. The entire model (including pretrained layers) are trained during our training with intra- and cross-modal losses. Further implementation, data pre-processing, and hyperparameter details are provided in supplementary material.

Metrics. Following standard practice [8, 17], we report the Area Under the ROC Curve (AUC) and Average Precision (AP) as performance metrics.

4.1 Main Results

Following the evaluation setup of [17, 40], we separately consider the five categories of FakeAVCeleb dataset (RVFA, FVRA-WL, FVFA-WL, FVFA-FS, FVFA-GAN). This allows us to gain insights into our performance under different types of deepfakes. Note that our methods achieve their performance *without seeing any of the identities or deepfake generations in the FakeAVCeleb dataset*, during its training. Our methods also do not require, nor has access to, any pristine (real) videos during inference.

SOTA Unsupervised Performance. In table 1, we observe that our combined method (Intra-Cross-modal) achieves a significant improvement over state-of-the-art unsupervised methods across all categories, and is only 2.7% lower than the best-performing *supervised* method [40] on average (AVG-FV). Note that the training datasets are not the same across different methods, so this shows an "as-is" comparison between methods. We train our method on VoxCeleb2 because LRS2 lacks identity labels needed for our intra-modal consistency training, and LRS3 has limited scenario diversity per identity. Nonetheless, we also report our method's results when trained on LRS3 in table 3 for completeness, and observe that it still outperforms prior unsupervised methods, albeit with slightly lower results than VoxCeleb2.

SOTA Performance on RVFA. For the detection of the challenging RVFA (real-video fake-audio) category, our performance (99.41% in AUC) surpasses all other supervised and unsupervised methods. Compared to the previous state-of-the-art unsupervised method, A-V Anomaly [17], our method achieves an 18.91% increase in AUC. Additionally, it outperforms the former top supervised method, AVFF [40], by 7.41%.

Complementary Benefits of Our Intra-Modal and Inter-Modal Methods. Another notable observation is that while our combined method (Intra-Cross-modal) achieves state-of-the-art performance, the performance of the two sub-scores (Intra-modal and Cross-modal) varies by the deepfake type and show complementary strengths (the worst AUC of the Cross-modal method is on FVFA-GAN, which is where the Intra-modal method has its best AUC). This observation provides further evidence for our theoretical justification that there is a trade-off between identity consistency and motion consistency, and therefore detecting both types of inconsistencies should be considered. We further discuss the separate performance of the Intra-modal and Cross-modal methods in the supplementary material C.3.

Table 2: AUC (%) comparison of our Intra-Cross-modal method on high/low quality compressed FakeAVCeleb dataset, and attacked KoDF dataset. The best and second-best results of supervised and unsupervised methods are in bold and underlined.

	Methods	FakeAVCeleb		KoDF
		HQ	LQ	Attacked
Supervised	Seferbekov[44]	**98.6**	**61.7**	**61.5**
	FTCN[51]	84.0	37.6	<u>58.3</u>
	LipForensics [31]	<u>97.6</u>	58.3	54.8
	Real Forensics[27]	88.3	52.9	55.5
	MDS-based FD[8]	64.7	<u>61.1</u>	55.4
	Joint AV[52]	55.1	55.2	47.1
Unsupervised	ICT [14]	68.2	66.9	61.5
	ICT-Ref[14]	71.9	71.2	78.0
	ID-Reveal [40]	70.2	70.8	73.4
	POI-Forensics [8]	<u>94.1</u>	**94.4**	<u>80.5</u>
	Ours (Intra-Cross-modal)	**95.5**	<u>85.9</u>	**84.2**$\pm$1.1

Generalization to Audio/Video Compression. Following the evaluation setting of [8], we construct a high quality and low quality version of the FakeAVCeleb dataset. In the High Quality (HQ) setting, the video is compressed using H.264 encoding with factor 23, and the audio is the same as original (44.1KHz). In the Low Quality (LQ) setting, the video is compressed using H.264 encoding with factor 40 and the audio is sampled with a sample rate of 16KHz. In table 2, we observe that our method outperforms all other unsupervised methods in the HQ compression setting (small video compression, no audio compression). In the LQ setting, our method is second to POI-Forensics, while still outperforming all other methods. However, it is important to note that POI-Forensics *requires* access to pristine videos during testing, which it can compress to the same LQ setting and directly compare the query videos against it. We could not include AV-Anomaly [17] in these evaluations due to lack of public inference and training code.

Generalization to KoDF. We also evaluated our method using the adversarially attacked subset of the KoDF [29] (both real and deepfake videos are adversarially attacked using the Fast Gradient Sign Method (FGSM) [18]). In this setting, POI-Forensics [8] randomly sampled 276 original videos and 544 synthesized videos as positive and negative samples for evaluation. The specific method for random selection of data was not disclosed by [8]. Therefore, for a robust and fair comparison, we repeat our evaluation 1000 times following their setting, each time with new random choice of 276 original and 544 synthesized videos from KoDF, which contains 5365 original and 15044 synthesized videos, and report the average and standard deviation of our method's AUC (%). In table 2, we observe that our method outperforms all supervised and unsupervised methods on this challenging task, where not only the language of the videos are completely different than training, but all the samples are also adversarially attacked. This empirical evidence further validates our theoretical insight (supplementary material C.1) that deepfakes contain inconsistencies in identity or motion.

Can Our Backbone Pre-trained Models Discover Inconsistencies Without Training? We utilize pretrained neural networks as the backbone for our identity, visual, and audio encoders that are originally trained for identity and speech recognition tasks. So it is possible that these models a priori can identify certain intra- and cross-modal inconsistencies. To verify the importance of our training process for discovering inconsistencies, we directly use our backbone pretrained models without further training to assess their inherent capability

Table 3: The effect of training dataset on the deepfake detection performance of our proposed methods on FakeAVCeleb. Without training with our objectives (in red), our backbone pertained visual, audio, and identity encoders have little success. Training with our objectives brings about significant performance enhancement, albeit slightly lower on LRS3 compared to VoxCeleb2.

	Method	Modality	Dataset	Category			
				RVFA		AVG-FV	
				AP	AUC	AP	AUC
Unsupervised	Ours (Intra)	V	None	-	-	63.62	63.58
	Ours (Intra)	V	LRS3	-	-	75.46	75.60
	Ours (Intra)	V	VoxCeleb2	-	-	76.15	77.86
	Ours (Cross)	AV	None	52.15	52.44	51.20	50.19
	Ours (Cross)	AV	LRS3	99.00	99.03	92.10	93.15
	Ours (Cross)	AV	VoxCeleb2	99.68	99.65	93.88	95.62
	Ours (Intra-Cross)	AV+V	None	52.50	53.07	63.61	63.64
	Ours (Intra-Cross)	AV+V	LRS3	97.79	98.70	94.50	95.08
	Ours (Intra-Cross)	AV+V	VoxCeleb2	**98.49**	**99.41**	**95.74**	**96.81**

in detecting inconsistencies, hence detecting deepfakes. In table 3, we observe that these pretrained models alone have almost no capability in detecting deepfakes through identifying intra-modal or cross-modal inconsistencies, suggesting that they fail to discern subtle variations in identity features and mismatches between audio and visual modalities. This highlights the effectiveness of our proposed loss functions and training strategies, which explicitly guide the models to recognize the intrinsic inconsistencies in deepfake videos, as discussed in section 3, thus enabling accurate deepfake detection.

5 Conclusion

In this work, we introduced a novel theoretical analysis that suggests motion or identity inconsistencies are inevitable in deepfake videos that transfer the motion of a source video to a target video. We also proposed unsupervised methods to explicitly detect and pinpoint identity and motion inconsistencies in deepfakes, and combined them into the Intra-Cross-modal method that achieved a new state-of-the-art performance on the FakeAVCeleb dataset. We also showed that our method generalizes to videos of other languages (KoDF Korean speech deepfake dataset), adversarial attacks, and small video compression. Furthermore, compared to existing deepfake detection methods, our method is more scalable because it does not require any pristine (real) videos during inference, generalizable because it only trains on real videos, reliable because it does not explicitly use likelihood estimation in high dimensions, and explainable because it explicitly discovers verifiable inconsistent segments in a video. Our findings reveal several directions for future research. First, directly measuring the terms in the proposed information-theoretical upper bound in the supplementary material C.1 for various deepfake methods can more directly verify the bound and be a new way of deepfake detection in itself. Second, the proposed Intra-modal method could be further improved by building inductive biases into the architecture that encourages attending to finer visual details. Finally, the proposed consistency losses could be applied to other video types beyond talking-heads, such as to multimodal full-body or aerial videos.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition. *CoRR*, abs/1809.00496, 2018. URL http://arxiv.org/abs/1809.00496.

[2] Shruti Agarwal, Hany Farid, Tarek El-Gaaly, and Ser-Nam Lim. Detecting deep-fake videos from appearance and behavior. In *2020 IEEE international workshop on information forensics and security (WIFS)*, pages 1–6. IEEE, 2020.

[3] Nicholas Carlini and Hany Farid. Evading deepfake-image detectors with white-and black-box attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 658–659, 2020.

[4] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. Not made for each other- audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 439–447, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413700. URL https://doi.org/10.1145/3394171.3413700.

[5] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. Not made for each other- audio-visual dissonance-based deepfake detection and localization, 2021.

[6] Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. *CoRR*, abs/1611.05358, 2016. URL http://arxiv.org/abs/1611.05358.

[7] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *CoRR*, abs/1806.05622, 2018. URL http://arxiv.org/abs/1806.05622.

[8] D. Cozzolino, A. Pianese, M. Niesner, and L. Verdoliva. Audio-visual person-of-interest deepfake detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 943–952, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society. doi: 10.1109/CVPRW59228.2023.00101. URL https://doi.ieeecomputersociety.org/10.1109/CVPRW59228.2023.00101.

[9] Davide Cozzolino and Luisa Verdoliva. Noiseprint: a cnn-based camera model fingerprint. *CoRR*, abs/1808.08396, 2018. URL http://arxiv.org/abs/1808.08396.

[10] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection, 2021.

[11] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019. doi: 10.1109/CVPR.2019.00482.

[12] Jiankang Deng, Jia Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and Song Shi. Lightweight face recognition challenge. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[13] Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Towards a robust deepfake detector: Common artifact deepfake detection model. *arXiv preprint arXiv:2210.14457*, 2022.

[14] Xiaoyi Dong, Jianmin Bao, Dongdong Chen, Ting Zhang, Weiming Zhang, Nenghai Yu, Dong Chen, Fang Wen, and Baining Guo. Protecting celebrities from deepfake with identity consistency transformer, 2022.

[15] Nikita Drobyshev, Jenya Chelishev, Taras Khakhulin, Aleksei Ivakhnenko, Victor Lempitsky, and Egor Zakharov. Megaportraits: One-shot megapixel neural head avatars. *arXiv preprint arXiv:2207.07621*, 2022.

[16] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883, June 2021.

[17] 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 (CVPR)*, pages 10491–10503, June 2023.

[18] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

[19] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

[20] 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, 2018. doi: 10.1109/AVSS.2018.8639163.

[21] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection, 2021.

[22] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Leveraging real talking faces via self-supervision for robust forgery detection, 2022.

[23] 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 (CVPR)*, June 2016.

[24] Yihao Huang, Felix Juefei-Xu, Run Wang, Qing Guo, Lei Ma, Xiaofei Xie, Jianwen Li, Weikai Miao, Yang Liu, and Geguang Pu. Fakepolisher: Making deepfakes more detection-evasive by shallow reconstruction. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1217–1226, 2020.

[25] 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.

[26] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:852–863, 2021.

[27] Hasam Khalid, Shahroz Tariq, and Simon S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *CoRR*, abs/2108.05080, 2021. URL https://arxiv.org/abs/2108.05080.

[28] Minchul Kim, Anil K. Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18750–18759, June 2022.

[29] Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset. *CoRR*, abs/2103.10094, 2021. URL https://arxiv.org/abs/2103.10094.

[30] Jiazhi Li, Mahyar Khayatkhoei, Jiageng Zhu, Hanchen Xie, Mohamed E Hussein, and Wael AbdAlmageed. Information-theoretic bounds on the removal of attribute-specific bias from neural networks. *arXiv preprint arXiv:2310.04955*, 2023.

[31] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection, 2020.

[32] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018.

[33] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*, 2018.

[34] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017.

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[36] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *European conference on computer vision*, pages 667–684. Springer, 2020.

[37] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2823–2832, 2020.

[38] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1xwNhCcYm.

[39] 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.

[40] 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 (CVPR)*, pages 27102–27112, June 2024.

[41] 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.

[42] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[45] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. *CoRR*, abs/1901.08971, 2019. URL http://arxiv.org/abs/1901.08971.

[46] Selim Seferbekov. Deepfake detection (dfdc) solution by @selimsef, 2020. URL https://github.com/selimsef/dfdc_deepfake_challenge.

[47] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[48] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.

[49] Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. Audio-driven talking face video generation with natural head pose. *CoRR*, abs/2002.10137, 2020. URL https://arxiv.org/abs/2002.10137.

[50] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[51] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection, 2021.

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

[53] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, and Jie Zhou. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. *CoRR*, abs/2103.04098, 2021. URL https://arxiv.org/abs/2103.04098.