Lips Are Lying: Spotting the Temporal Inconsistency between Audio and Visual in Lip-Syncing DeepFakes

Weifeng Liu, Tianyi She,
Jiawei Liu, Boheng Li,
Dongyu Yao, Ziyou Liang,
Run Wang*

NEURAL INFORMATION PROCESSING SYSTEMS

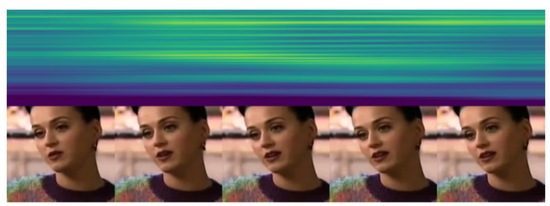
Abstract

DeepFake technology has advanced in high-quality video synthesis but poses significant security risks. Lip-forgery videos, which don't alter identity or show visual artifacts, challenge existing detection methods, often causing them to fail. We introduce a novel approach for identifying them by detecting inconsistencies between lip movements and audio. Our method achieves over 95.3% accuracy in detecting LipSync videos and up to 90.2% in real-world scenarios. Our main contributions can be summarized as follows:

- We propose LipFD, the first method dedicated to LipSync detection.
- We unveil a key insight that exploits the discrepancies between lip movements and audio signals for fine-grained forgery detection.
- We construct the first large scale audio-visual LipSync dataset AVLips, with up to 340,000 samples

Consistency between lip movement and audio

(a) shows the correlation between lip movements and corresponding spectrogram in genuine pattern. When the woman starts talking, the middle and high frequencies in the spectrum are lighted. Over time, the energy gradually fades and shifts from middle to lower frequencies. (b) the first two frames show a highlighted high-frequency spectrum, contradicting the man not speaking. In the third frame, an unexpected lip opening appears at the darkest part of the spectrum. The mouth cannot change so drastically within a single frame, and this lip shape contradicts the spectrum information.

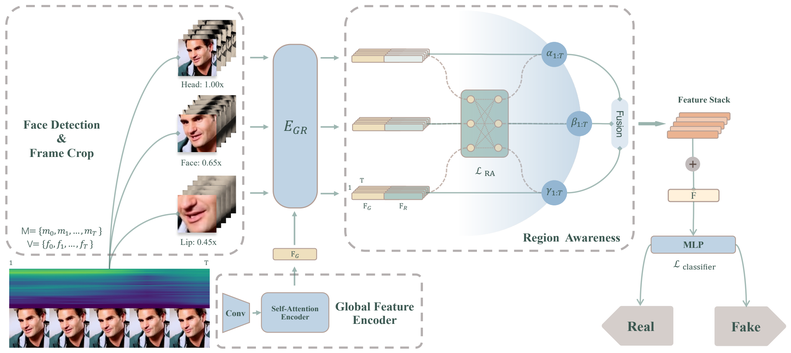


(a) real pattern

(b) fake pattern

* Corresponding author. Email: wangrun@whu.edu.cn

Overview of LipFD framework



Pipeline of our model.

Blue components represent our main modules in LipFD. The input image was generated by pre-processing, which consists of T frames in the target video and their audio spectrogram. (a) The aim of Global Feature Encoder, a self-attention model, is to extract long-term information between video frames and audio, finding unreasonable correspondences between lip movements and audio. (b) E_{GR} encodes three series of crops, focusing on different parts for each region, and concatenates them with global feature F_G. (c) The Region Awareness module assigns corresponding weights to the features based on their importance. (d) All features are fused together into a unified representation F based on their respective weights for final inference.

Experiments

Table 1: **Cross-datasets validation.** Results on AVLips, FF++, and DFDC are reported, including acc, ap, fpr and fnr. The best result is highlighted in bold, while the second-ranking one is underscored. Throughout the entire experiment, the threshold for the AP metric was set to 0.5.

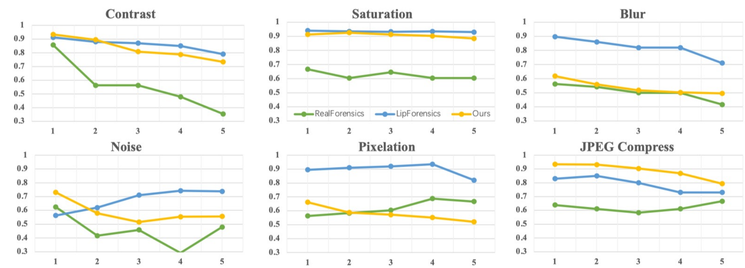
Method	AVLips				FF++				DFDC			
	ACC	AP	FPR	FNR	ACC	AP	FPR	FNR	ACC	AP	FPR	FNR
CViT	65.54	56.68	0.07	0.61	62.86	54.17	0.24	0.50	70.99	58.06	0.06	0.50
DoubleStream	75.52	67.72	0.13	0.36	91.02	87.64	**0.03**	0.14	77.39	69.28	0.21	0.24
UniversalFakeDetect	50.03	50.02	0.99	**0.01**	50.43	50.16	0.99	***0.01***	49.86	49.94	0.98	**0.01**
SelfBlendedImages	49.99	52.13	0.07	0.51	64.59	57.93	0.17	0.53	48.47	49.06	0.15	0.50
RealForensics	91.78	90.14	**0.02**	0.14	93.57	91.32	**0.03**	0.10	92.54	**91.62**	**0.00**	0.15
LipForensics	86.13	81.56	0.18	0.10	94.03	**93.25**	0.04	0.08	90.75	87.32	0.08	0.11
LipFD (Ours)	**95.27**	**93.08**	0.04	0.04	**95.10**	76.98	0.06	0.05	**94.53**	78.61	0.08	0.04

Table 2: **Cross-manipulation generalisation.** Evaluation scores when videos are exposed to various unseen forgery algorithms.

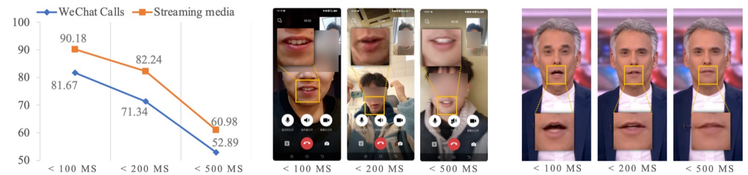
Method	ACC	FPR	FNR	AUC
Wav2Lip (Dynamic)	95.27	0.04	0.04	95.27
MakeItTalk (Static)	**96.93**	**0.02**	**0.03**	**96.89**
TalkLip (Dynamic)	79.33	0.34	0.04	80.36

Table 3: **Overall ablation results regarding core modules.** We evaluated our model's performance after removing components listed in the left column.

Component	ACC	AP	FPR	FNR	AUC
Global Encoder	95.07	91.81	0.02	0.07	95.09
Global-Region Encoder	72.52	64.38	**0.01**	0.53	72.50
Region Awareness	76.45	72.65	0.38	0.09	76.32
Full model	**95.27**	**93.08**	0.04	**0.04**	**95.27**



Robustness against various unseen corruptions. Average AUC scores across five intensity levels for various corruptions.



Performance in real scenarios. The x-axis represents network delay time, where a higher delay indicates a degradation in image transmission quality and clarity. Consequently, this degradation adversely impacts the audio-video synchronization in WeChat video calls.

AVLips: A high-quality audio-visual dataset for LipSync detection

To the best of our knowledge, the majority of public DeepFake datasets consist solely of videos or images, with no specialized one specifically dedicated to LipSync detection available. To fill this gap, we construct a high-quality Audio-Visual Lip-syncing Dataset, AVLips, which contains up to 340,000 audio-visual samples generated by several SOTA LipSync methods.