Supplemental Material for
Audio-Visual Person-of-Interest DeepFake Detection

Davide Cozzolino[1] Alessandro Pianese[1] Matthias Nießner[2] Luisa Verdoliva[1]
[1]University Federico II of Naples [2]Technical University of Munich

In this supplemental document, we present additional ablation studies (Sec.1) and more results to prove the robustness capability of our method (Sec.2). Moreover, we briefly describe the state of the art methods we compare to, (Sec.3) and give some more details about the dataset used in the experiments (Sec.4). Finally, we summarize the limitations of our method (Sec.5).

1. Additional ablation studies

In this Section, we conduct additional experiments to show that our approach outperforms some state-of-the-art audio-visual speaker verification methods for both the person identification task and the deepfake detection task. We consider three reference methods, all relying on a contrastive learning paradigm:

- In SyncNet[1] [17], a robust speaker identity representation is proposed, based on an end-to-end self-supervised approach. Specific constraints are imposed on the identity, that should change slowly over time, on the content, that instead should change quickly, and on both factors that are enforced to be represented independently of one-another through a disentangling constraint.
- The approach proposed in [23] learns an audio-visual embedding for person verification by using two separate networks for audio and video that are jointly trained. Then, fusion is performed at feature-level, by relying on an attention mechanism that learns the salient modality of input data.
- In [21] an audio-visual fusion system is also proposed where, after concatenating the two features, fusion is performed by means of a multilayer perceptron.

First, we study the person identification problem using the VoxCeleb2 dataset [4] and considering from 10 to 200 different identities (not included in training). For training, we collect 100 video segments for each subject. At testing time, we use 1-Nearest Neighbor classification. Performance is evaluated in terms of accuracy on 10 video segments for each subject. Results are shown in Table 1. Since our method and [17] are able to provide results even using a single modality we include also these results, in the upper part of the table. Then, in the lower part we report fusion-

[1]available at https://github.com/joonson/syncnet_trainer

		10	50	100	150	200
Random Clas.		10.0	2.0	1.0	0.7	0.5
audio	[17]	69.0	43.2	31.7	25.8	22.4
	ours	81.0	67.6	61.1	54.3	50.1
video	[17]	53.0	32.8	24.8	20.6	18.4
	ours	75.0	62.4	57.8	53.5	50.0
both	[17]	82.0	56.6	45.0	39.9	35.9
	[21]	**92.0**	75.8	69.6	65.2	63.0
	[23]	90.0	79.0	72.8	69.8	66.4
	ours	91.0	**82.0**	**77.9**	**76.4**	**73.3**

Table 1. Results in term of ACC (%) on the person identification task considering a variable number of identities, from 10 to 200.

	v	a	ai	[17]	[21]	[23]	Ours
AUC (%)	✓			66.3	61.0	60.8	71.4
	✓		✓	80.5	93.4	94.7	96.0
		✓	✓	91.0	73.6	80.4	90.9
	✓	✓	✓	93.8	93.7	95.1	97.5
	AVG			82.9	80.4	82.8	**88.9**
Pd@10% (%)	✓			19.0	26.3	23.3	37.3
	✓		✓	47.9	86.8	88.2	94.5
		✓	✓	57.8	53.1	51.7	76.6
	✓	✓	✓	77.0	90.4	90.4	95.6
	AVG			50.4	64.1	63.4	**76.0**

Table 2. Results in terms of AUC and Pd%. We compare our approach with other different strategies of audio-visual POI identification considering several scenarios, i.e. four situations identified by the checkmarks in the first three columns, indicating a video manipulation (v), an audio manipulation (a) and an audio inconsistency (ai).

	No noise	video-only noise	audio-only noise	both noise
AUC	89.9	78.8	75.4	74.2
Pd@10%	74.3	50.3	55.5	52.7

Table 3. Results of our method on the noisy KoDF subset. We added noise on video and audio and evaluated the performance considering video-only, audio-only and both.

based results for all methods, where both audio and video are exploited. In all cases, our method outperforms all references, with the only exception of fusion when only 10 identities are involved (where it is second best). The performance gain is especially significant in the most challenging cases where a large number of identities are considered. As an example, with 200 identities, the proposed method has an accuracy of 73.3%, 7 points better than the second best.

For deepfake detection, results are reported in Table 2 in a similar way as in the main paper (see Section 4.2, Table 1). We identify four groups according to the checkmarks in the first three columns, indicating video manipulation (v), audio manipulation (a), audio inconsistency (ai). The dataset used for the analysis is again a subset of FakeAVCelebV2 and KoDF, comprising a total of more than 140 subjects. To ensure a fair comparison, reference methods have been trained on the same dataset used for our approach. Moreover, for the methods proposed in [21,23] we used our backbone. Also in this scenario our approach ensures a clearly superior performance with respect to all references both in terms of AUC, with an average improvement of 6 percent points with respect to the second best, and in terms of $\mathrm{Pd@10\%}$, in which case the average improvement reaches 12 percent points.

2. Additional robustness analysis

In this Section, we provide some more insights on the robustness of our approach. In the main paper, we analyzed how compression and adversarial attacks impair the performance of our detector (Section 4.3, Table 2, and Table 3). Here, we focus on Gaussian noise addition which is well-known to strongly impair deepfake detection performance when only the video has been modified [10]. Moreover, noise is a serious issue also for audio-based speaker verification [12].

In our experiments, we consider again a subset of the KoDF dataset [16] and add Gaussian noise to both audio and video signals, with random intensity amounting to an SNR going from 5 to 15 dB for the audio signal, and a PSNR going from 13 to 27 dB for the video signal. In Table 3, we show the impact of noise addition on our detector. We consider separately the cases where only the video, only the audio or both modalities are attacked. In any case, the detector uses both modalities. As expected, for such intense levels of noise, the performance reduces sharply. Nonetheless, even in this scenario our method keeps ensuring a reasonable performance, with an AUC of around 74% when both audio and video are attacked.

This consideration is further reinforced by comparing results with state-of-the-art detectors (Table 4). To ensure a fair comparison, we run our method using only the video signal, as all references do. Even so, a large performance improvement is observed with respect to all competitors, with over 8% AUC gain and 16% $\mathrm{Pd@10\%}$ gain with respect to the second best, ICT-Ref.

3. State-of-the-art methods

In the main paper, we compare our proposal with six state-of-the-art approaches:

Seferbekov [22] is the first-place solution of the Kaggle Deepfake Detection Challenge [6]. It is based on an ensemble of seven Efficientnet-B7 models trained with strong augmentation and a frame-by-frame analysis.

FTCN (Fully Temporal Convolution Network) [26], it focuses on temporal cues, exploiting short-term flickering with a Fully Temporal Convolution Network and long-term incoherence with a Temporal Transformer.

LipForensics [10] uses a spatio-temporal network, pretrained to perform visual speech recognition (lipreading), to detect semantic irregularities in the mouth movements.

Real Forensics [9] is a teacher-student network that uses the audio-video pair in a multitask fashion. The student network performs both real/fake classification and features extraction. The teacher is used to provide the target features.

MDS-based FD [3] is an audio-visual fake detector (FD)

	Seferbekov	FTCN	LipFor.	Real.For.	MDS-based FD	Joint AV	ICT	ICT-Ref	ID-Reveal	ours (video)
AUC	68.4	50.2	54.5	62.5	69.1	50.4	59.9	72.6	62.8	**79.1**
Pd@10%	31.3	10.1	15.2	16.2	22.2	12.3	23.9	42.8	16.2	**49.9**

Table 4. Results on the noisy KoDF subset. We compare our approach with Seferbekov [22], FTCN (Fully Temporal Convolution Network) [26], LipForensics [10], RealForensics [9], MDS-based FD [3], ICT, ICT-Ref [8], and ID-Reveal [5].

based on modality dissonance score (MDS), a similarity measure between audio and visual streams. The idea is to capture inconsistencies such as lack of lip synchronization, unnatural facial movements or asymmetries.

Joint AV [27] is a multimodal detector that handles video and audio streams separately with their own labels. On top of this, the model also synchronizes the representations from the two streams to discriminate synchronization patterns between pristine and manipulated data.

ICT (Identity Consistency Transformer) [8]. It focuses on finding identity-based inconsistencies between the inner region and outer one of the face using a transformer-based architecture. A reference-assisted variant (**ICT-Ref**) is also proposed, that assumes the availability of a reference set of real videos.

ID-Reveal [5] is a single-modality (visual information only) identity-based detector. It relies on 3D Morphable Models and an adversarial game to improve the discrimination performance.

4. Deepfake video datasets

In this Section we briefly describe the four deepfake video datasets used in our experiments:

pDFDC, preview DeepFake Detection Challenge dataset [7]. We show results on 44 individuals which have more than 9 videos with a total of 920 real videos and 2925 fake ones.

FakeAVCelebV2, Audio-Video Deepfake dataset [11]. It comprises 500 real videos coming from Voxceleb2 [4] and about 20,000 fake videos generated by both face-swapping (Faceswap [15], Faceswap GAN (FSGAN) [18]) and facial reenactment (Wav2Lip [20]) methods. Fake audios are generated by a transfer learning-based real-time voice cloning tool (SV2TTS [13]). These methods are then used individually or combined together giving rise to five categories of manipulated videos.

KoDF, a large-scale Korean DeepFake dataset [16]. It includs three face swapping manipulations: FaceSwap [1], DeepFaceLab [19] and FSGAN [18] and three face-reenactment ones: First Order Motion Model (FOMM) [24], Audio-driven face synthesis ATFHP [25] and Wav2Lip [20]. We consider a test-set comprising 276 real videos and 544 fake ones.

DF-TIMIT, DeepFake-TIMIT [14]. An open source GAN-based face swapping method is used [2] with two different input dimensions of the GAN network so as to obtain two manipulated videos for each real one. We report results for videos of at least 4 seconds, for a total of 290 real videos and 580 fake ones.

Figure 1. Examples of videos on which our approach fails. All videos come from the pDFDC dataset [7].

5. Limitations

While our method does not need to include any fake videos in training, it requires a large dataset of audio-visual information from several different subjects for training and it needs a few (around 10) reference pristine videos of the target subject at testing time. We also noticed a drop in performance on faces seen in profile (Fig.1). This is probably due to the preponderance of frontal-pose videos in the training set and could probably be solved by a better balancing of training samples or by suitable forms of augmentation.

References

[1] Faceswap. https://github.com/deepfakes/faceswap. 3

[2] FaceSwap-GAN. https://github.com/shaoanlu/faceswap-GAN. 3

[3] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian. Not made for each other- audio-visual dissonance-based deepfake detection and localization. In *ACM International Conference on Multimedia*, 2020. 2

[4] J.S. Chung, A. Nagrani, and A. Zisserman. VoxCeleb2: Deep speaker recognition. In *Interspeech*, 2018. 1, 3

[5] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, and L. Verdoliva. ID-Reveal: Identity-aware DeepFake Video Detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[6] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton Ferrer. The deepfake detection challenge dataset. *arXiv preprint arXiv:2006.07397*, 2020. 2

[7] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton Ferrer. The deepfake detection challenge (DFDC) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019. 3

[8] X. Dong, J. Bao, D. Chen, T. Zhang, W. Zhang, N. Yu, D. Chen, F. Wen, and B. Guo. Protecting celebrities from deepfake with identity consistency transformer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[9] A. Haliassos, R. Mira, S. Petridis, and M. Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[10] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[11] K. Hasam, T. Shahroz, K. Minha, and S.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. 3

[12] H.S. Heo, B.-J. Lee, J. Huh, and J.S. Chung. Clova baseline system for the voxceleb speaker recognition challenge 2020. *arXiv preprint arXiv:2009.14153*, 2020. 2

[13] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 3

[14] P. Korshunov and S. Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 3

[15] I. Korshunova, W. Shi, J. Dambre, and L. Theis. Fast face-swap using convolutional neural networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[16] P. Kwon, J. You, G. Nam, S. Park, and G. Chae. KoDF: A large-scale korean deepfake detection dataset. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[17] A. Nagrani, J.S. Chung, S. Albanie, and A. Zisserman. Disentangled speech embeddings using cross-modal self-supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020. 1

[18] Y. Nirkin, Y. Keller, and T. Hassner. FSGAN: Subject agnostic face swapping and reenactment. In *IEEE International Conference on Computer Vision (ICCV)*, 2019. 3

[19] I. Petrov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Um'e, Mr. dpfks, RP Luis, J. Jiang, S. Zhang, P. Wu, B. Zhou, and W. Zhang. DeepFaceLab: A simple, flexible and extensible face swapping framework. *arXiv preprint arXiv:2005.05535*, 2020. 3

[20] K.R. Prajwal, R. Mukhopadhyay, V.P. Namboodiri, and C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM International Conference on Multimedia*, 2020. 3

[21] L. Sari, K. Singh, J. Zhou, L. Torresani, N. Singhal, and Y. Saraf. A Multi-view approach to audio-visual speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021. 1, 2

[22] S. Seferbekov. *DeepFake Detection (DFDC) Team Sefer.* https://github.com/selimsef/dfdc_deepfake_challenge. 2

[23] S. Shon, T.-H. Oh, and J. Glass. Noise-tolerant audio-visual online person verification using an attention-based neural network fusion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019. 1, 2

[24] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[25] R. Yi, Z. Ye, J. Zhang, H. Bao, and Y.-J. Liu. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*, 2020. 3

[26] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen. Exploring temporal coherence for more general video face forgery detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[27] Y. Zhou and S.-N. Lim. Joint audio-visual deepfake detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 3