AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection

Supplementary Material

A. Overview

This document is structured as follows:

- Sec. B: Implementation details
- Sec. C: Dataset details
- Sec. D: Additional results
- Sec. E: Visual examples of challenging scenarios

B. Implementation Details

B.1. Representation Learning Stage

Inputs. We draw samples from the LRS3 dataset [1], which exclusively contains real videos. We preprocess all videos as explained in Sec. 3.1 in the main paper. The audio stream is converted to a Mel-spectrogram of 128 Mel-frequency bins, with a 16 ms Hamming window every 4 ms. We randomly sample video clips of $T = 3.2$s in duration, sampling 16 visual frames and 768 audio frames (Mel) with clipping/padding where necessary. The 16 visual frames are uniformly sampled such that they are at the first and third quartile of a temporal slice (2 frames/slice $\times$ 8 slices). The visual frames are resized to 224×224 spatially and are augmented using random grayscaling and horizontal flipping, each with a probability of 0.5. We make sure that in a given batch, for each sample we draw another sample from the same video but at a different time interval to make sure the model is exposed to the notion of temporal shifts when computing the contrastive loss. Both audio and visual modalities are normalized.

Architecture. We adopt the encoder and decoder architectures of each modality from the VideoMAE [29] based on ViT-B [5]. Each of the A2V/V2A networks is composed of a linear layer to match the number of tokens of the other modality followed by a single transformer block.

Optimization. We initialize the audio encoder and decoder using the checkpoint of AudioMAE [12] pretrained on AudioSet-2M [7] and the visual encoder and decoder using the checkpoint of MARLIN [2] pretrained on the YouTubeFace [30] dataset. Subsequently, we train the representation learning framework end-to-end, using the AdamW optimizer [21] with a learning rate of 1.5e-4 with a cosine decay [20]. The weights of the losses are as follows: $\lambda_c = 0.01$, $\lambda_{rec} = 1.0$, and $\lambda_{adv} = 0.1$, which were chosen empirically and based on previous research [2, 8]. We train for 500 epochs with a linear warmup for 40 epochs using a batch size of 32 and a gradient accumulation interval of 2. The training was performed on 4 RTX A6000 GPUs for approximately 60 hours.

Method	Modality	DF-TIMIT		DFDC	
		AP	AUC	AP	AUC
Xception [26]	V	86.0	90.5	68.0	67.6
LipForensics [9]	V	96.7	98.4	76.8	77.4
FTCN [33]	V	**100.**	99.8	70.5	71.1
RealForensics [10]	V	99.2	99.5	<u>82.9</u>	<u>83.7</u>
AVFF (Ours)	AV	**100.**	**100.**	**87.0**	**86.2**

Table 1. **Cross-Dataset Generalization.** We evaluate our model's performance against baselines by testing the model trained on the FakeAVCeleb dataset, on the DF-TIMIT dataset and a subset of the DFDC dataset. Best result is in bold, and second best is underlined.

B.2. Deepfake Classification Stage

Inputs. We draw samples from FakeAVCeleb [16], which consists of deepfake videos where either or both audio and visual modalities have been manipulated. The preprocessing and sampling strategy is similar to that of Stage 1, except we do not draw an additional sample from the same video clip as we do not use a contrastive learning objective at this stage. We employ weighted sampling to mitigate the issue of class imbalance between real and fake samples.

Architecture. Each of the uni-modal patch reduction networks is a 3-layer MLP, while the classifier head is a 4-layer MLP. We do not make any changes to the representation learning architecture.

Optimization. We initialize the representation learning framework using the pretrained checkpoint obtained from Stage 1. Subsequently, we train the pipeline end-to-end, using the AdamW optimizer [21] with a cosine annealing with warm restarts scheduler [20] with a maximum learning rate of 1.0e-4 for 50 epochs with a batch size of 32. The training was performed on 4 RTX A6000 GPUs for approximately 10 hours.

C. Dataset Details

LRS3 [1]. This dataset introduced by Afouras *et al*. exclusively comprises of real videos. It consists of 5594 videos spanning over 400 hours of TED and TED-X talks in English. The videos in the dataset are processed such that each frame contains faces and the audio and visual streams are in sync.

FakeAVCeleb [16]. The FakeAVCeleb dataset is a deepfake detection dataset, which consists of 20,000 video clips in total. It comprises of 500 real videos sampled from the VoxCeleb2 [3] and 19500 deepfake samples generated using different manipulation methods applied on the set of real

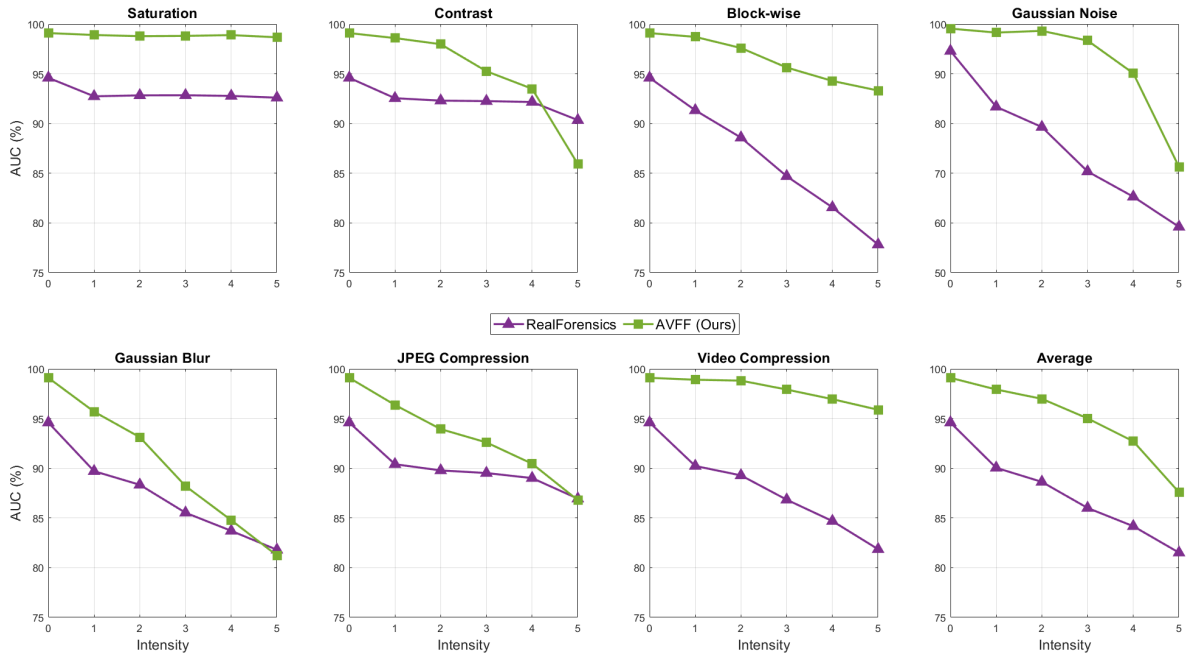

Figure 1. **Robustness to Unseen Visual Perturbations**. We illustrate AUC scores (%) as a function of different levels of intensities for various visual perturbations evaluated on the test set of FakeAVCeleb. Our model is more robust than RealForensics [10], which is the current state-of-the-art in robustness to unseen visual perturbations.

videos. The dataset consists of the following manipulations where the deepfake algorithms used in each category are indicated within brackets.

- RVFA: Real Visuals - Fake Audio (SV2TTS [13])
- FVRA-FS: Fake Visuals - Real Audio (FaceSwap [18])
- FVFA-FS: Fake Visuals - Fake Audio (SV2TTS + FaceSwap)
- FVFA-GAN: Fake Visuals - Fake Audio (SV2TTS + FaceSwapGAN[22])
- FVRA-GAN: Fake Visuals - Real Audio (FaceSwapGAN)
- FVRA-WL: Fake Visuals - Real Audio (Wav2Lip [25])
- FVFA-WL: Fake Visuals - Fake Audio (SV2TTS + Wav2Lip)

KoDF [19]. This dataset is a large-scale dataset comprising real and synthetic videos of 400+ subjects speaking Korean. KoDF consists of 62K+ real videos and 175K+ fake videos synthesized using the following six algorithms: FaceSwap [18], DeepFaceLab [23], FaceSwapGAN[22], FOMM [28], ATFHP [31], and Wav2Lip [25]. We use a subset of this dataset following [6] to evaluate the cross-dataset generalization performance of our model (Tab. 3 in the main paper).

DF-TIMIT [17]. The Deepfake TIMIT dataset comprises deepfake videos manipulated using FaceSwapGAN [22]. The real videos used for manipulation have been sourced by sampling similar-looking identities from the VidTIMIT [27] dataset. We use their higher-quality (HQ) version, which consists of 320 videos, in evaluating cross-dataset generalization performance.

DFDC [4]. The DeepFake Detection Challenge (DFDC) dataset is another deepfake dataset that consists of samples with fake audio besides FakeAVCeleb. It consists of over 100K video clips in total generated using deepfake algorithms such as MM/NN Face Swap [11], NTH [32], FaceSwapGAN [22], StyleGAN [15], and TTS Skins [24]. We use a subset of this dataset consisting of 3215 videos, as used in [9, 10] to evaluate the model's cross-dataset generalization performance.

D. Additional Results

In extending our analysis beyond the results outlined in Sec. 4 of the main paper, we conducted additional experiments to provide a more comprehensive evaluation of our model's performance. This supplementary investigation aims to enhance our understanding and confidence in the efficacy of the proposed approach.

D.1. Cross-Dataset Generalization

In addition to the cross-dataset generalization evaluation reported on the KoDF dataset (Tab. 3 of the main paper), we further evaluate cross-dataset generalization on the DF-TIMIT dataset and a subset of the DFDC dataset following [9, 10]. We are limited to comparing against baselines with open-source codes. We were unable to obtain the models nor

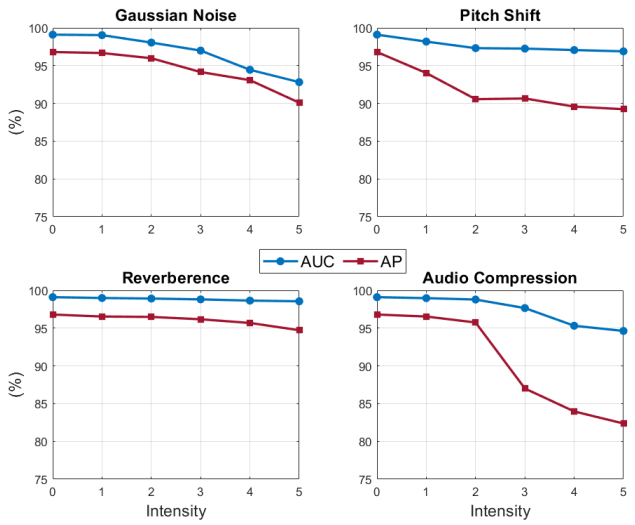

Figure 2. **Robustness to Unseen Audio Perturbations**. We illustrate the variation of AUC and AP scores (%) as a function of different levels of intensities for various audio perturbations evaluated on the test set of FakeAVCeleb. Overall our model depicts impressive robustness to audio perturbations.

results for the other baselines from the authors. As illustrated in Tab. 1, we achieve the best cross-dataset generalization performance, when evaluated on both DF-TIMIT and DFDC.

D.2. Robustness to Unseen Perturbations

In real-world scenarios, videos undergo post-processing (*e.g.* when sharing through social media platforms), which perturbs both audio and visual modalities. Hence, it is crucial for a model to be robust to unseen perturbations. To this end, we evaluate the performance of our model (trained without augmentations) on several unseen perturbations applied to each modality.

Visual Perturbations. Following [6, 9, 10], we evaluate the performance on the following perturbations: saturation, contrast, block-wise distortion, Gaussian noise, Gaussian blur, JPEG compression, and video compression on five different levels of intensities. The implementations for the perturbations and the levels of intensities were sourced from the official repository of DeeperForensics-1.0 [14]. We compare our model's performance against RealForensics [10], which has the current state-of-the-art performance in robustness to unseen visual perturbations [6, 10]. As depicted in Fig. 1, our model demonstrates enhanced robustness against unseen visual perturbations compared to RealForensics in most scenarios. Particularly, noteworthy improvements are observed in cases of block-wise distortion, Gaussian noise, and video compression.

Audio Perturbations. In this experiment, we subject the audio stream to a range of perturbations: Gaussian Noise, pitch shift, changes in reverberance, and audio compression.

Perturbation	Intensity Level				
	1	2	3	4	5
Gaussian Noise (SNR)	40	30	20	15	10
Pitch Shift (steps)	±2	±4	±6	±8	±10
Reverberance	20	40	60	80	100
Audio Compression (bitrate)	320k	256k	192k	128k	64k

Table 2. Parameters used to generate samples with audio perturbations at different levels of intensities.

	Method	ACC	AUC
(i)	Ours with Frozen Stage 1 + MLP	94.8	85.3
(ii)	Ours with Frozen Stage 1 + SVM	96.3	88.9
	AVFF (Ours)	**98.6**	**99.1**

Table 3. **Classification Performance on the Learned Representation.** We evaluate the classification performance of the learned representation by freezing the encoders and A2V/V2A networks and training only the downstream networks. We employ (i) an MLP similar to the proposed method, and (ii) a kernel SVM (RBF), as the classifier. Both classifiers yield reasonably high metrics, indicating the effectiveness of the learned representation at the end of Stage 1 in distinguishing between real and fake videos.

The performance of our model under these perturbations is illustrated across five intensity levels in Fig. 2. To generate the perturbed audio samples, we employ the following Python libraries: *torchaudio* (Gaussian noise, pitch shift), *pysndfx* (reverberance), and *pydub* (audio compression). The parameters used to generate samples at each intensity level are tabulated in Tab. 2. As seen in Fig. 2, overall our model is robust to various audio perturbations. A slight decrease in performance is seen in cases of Gaussian noise and pitch shift, with the increase in intensity. Notably, the model showcases high robustness to changes in reverberance, with minimal fluctuations across all intensity levels. However, a noticeable reduction in average precision is observed for high-intensity levels of audio compression, potentially due to artifacts introduced by the reduced bitrate in extreme compression scenarios.

D.3. Classification Performance on the Learned Representation

We further evaluate the learned representation at the end of Stage 1, by performing the downstream deepfake classification task with the weights of the encoders and the A2V/V2A networks frozen. We train two classifiers for the downstream task: (i) the classifier network described in the main paper, and (ii) kernel SVM using an RBF kernel (gamma=0.1, C=1.0). The results are reported in Tab. 3. Both classifiers yield reasonably high accuracy and AUC values. This is indicative of the highly discriminative nature of the learned representation at the end of Stage 1. This reinforces the analysis on the learned representation at the end of Stage 1,

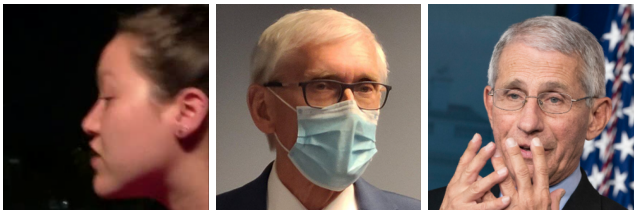

Figure 3. **Visual Examples of a Few Challenging Scenarios**. Images from left to right depict examples of extreme poses (*e.g.* near profile), occlusions with masks, and occlusions with hands across the face, which makes it challenging for our model to establish correspondence between the audio and visual modalities.

as discussed in Sec. 4.2 of the main paper.

E. Visual Examples of Challenging Scenarios

As discussed in Sec. 6 of the main paper, since we rely on audio-visual correspondence to distinguish between real and fake videos, scenarios where such correspondence cannot be established would be challenging. In Fig. 3, we depict a few such visual examples.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. 1

[2] 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. 1

[3] J Chung, A Nagrani, and A Zisserman. Voxceleb2: Deep speaker recognition. *Interspeech 2018*, 2018. 1

[4] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*, 2020. 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 1

[6] 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. 2, 3

[7] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 1

[8] Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. Contrastive audio-visual masked autoencoder. In *ICLR*. OpenReview.net, 2023. 1

[9] 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. 1, 2, 3

[10] 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. 1, 2, 3

[11] Dong Huang and Fernando De La Torre. Facial action transfer with personalized bilinear regression. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II 12*, pages 144–158. Springer, 2012. 2

[12] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. Masked autoencoders that listen. *Advances in Neural Information Processing Systems*, 35:28708–28720, 2022. 1

[13] 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. 2

[14] 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. 3

[15] 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. 2

[16] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021. 1

[17] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 2

[18] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3677–3685, 2017. 2

[19] 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. 2

[20] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2016. 1

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 1

[22] 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. 2

[23] 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. 2

[24] Adam Polyak, Lior Wolf, and Yaniv Taigman. Tts skins: Speaker conversion via asr. *arXiv preprint arXiv:1904.08983*, 2019. 2

[25] 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. 2

[26] 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. 1

[27] Conrad Sanderson and Brian C Lovell. Multi-region probabilistic histograms for robust and scalable identity inference. In *Advances in biometrics: Third international conference, ICB 2009, alghero, italy, june 2-5, 2009. Proceedings 3*, pages 199–208. Springer, 2009. 2

[28] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019. 2

[29] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 1

[30] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534. IEEE, 2011. 1

[31] 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. 2

[32] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019. 2

[33] 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. 1