Leveraging Real Talking Faces via Self-Supervision for Robust Forgery Detection - Extended Abstract

Alexandros Haliassos[1] Rodrigo Mira[1] Stavros Petridis[1,2] Maja Pantic[1,2]

[1]Imperial College London [2]Meta AI

{alexandros.haliassos14,rs2517,stavros.petridis04,m.pantic}@imperial.ac.uk

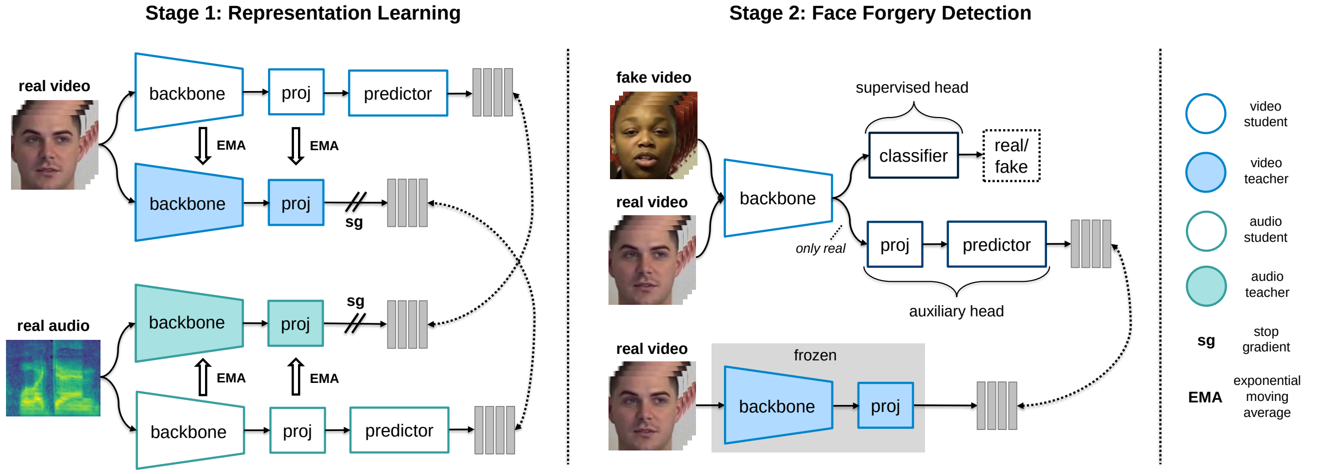

Figure 1. **The two stages of RealForensics**. In stage 1, the aim is to learn, in a self-supervised manner, frame-wise representations that capture information on natural facial behaviour and appearance. We utilise an audiovisual, cross-modal, student-teacher framework, whereby the student networks ingest real video and audio and try to predict the corresponding targets generated from the other modality. The teacher networks are momentum encoders that are updated via an exponential moving average (EMA), as in [11]. In stage 2, the detector performs face forgery classification, while predicting the video targets produced by the (now frozen) video teacher from stage 1; only real videos contribute to the prediction loss. The video student from stage 1 is used to initialise the backbone.

1. Introduction

Automatic face manipulation methods can realistically change someone's appearance or expression without requiring substantial human expertise or effort [8, 16, 18, 21, 27]. This technology's potential social harm has spurred considerable research efforts to detect forgery content [1, 4, 7, 10, 12–14, 17, 19, 24, 26, 31, 33, 35].

Although deep learning-based detectors can achieve high accuracy on in-distribution data, performance often plummets on videos generated using novel manipulation methods [3, 6, 14, 19, 21, 30, 35]. Various methods have been proposed to tackle cross-manipulation generalisation, including using data augmentation [30], truncating classifiers [3], and targeting the blending boundary in fake videos [19]. However, many still underperform on novel forgery types or focus on low-level cues easily corrupted by operations like compression [14]. Targeting both generalisation and robustness to corruptions, LipForensics [14] pre-trains on a large-scale lipreading dataset to focus on high-level inconsistencies in mouth movements, but it requires costly text transcriptions, limiting its scalability.

In this work, we are motivated by the observation that fake videos often exhibit anomalous facial movements and expressions, as well as subtle changes in facial form over time. Such cues are high-level in nature and thus more resilient to low-level corruptions. To target such cues, we propose a two-stage approach, termed *RealForensics*. In the first stage, partly inspired by BYOL [11], we propose to use a self-supervised, student-teacher framework to exploit the correspondence (in terms of *e.g.*, lexical content, emotion, identity) between the visual and auditory modalities in natural videos of talking faces. In the second stage, the forgery detector is tasked with performing classification while simultaneously predicting representations learned in the first stage, alleviating overfitting. Our experiments demonstrate state-of-the-art performance in cross-manipulation generalisation as well as high robustness to common corruptions.

2. Method

Our two-stage approach is depicted in Figure 1.

Stage 1: representation learning. Given a dataset of real videos and the corresponding audio, represented as log-mel spectrograms, we aim to learn representations that capture information associated with facial appearance and behaviour. Cues like facial movements are temporally fine-grained, and hence we learn *temporally dense* representations, *i.e.*, an embedding per frame. In this work, we use the LRW dataset [5].

Our architecture consists of a student and teacher pair for each modality. The teachers produce targets for the students from the other modality to predict. Specifically, backbone networks produce embeddings passed through linear projectors to yield dense targets. The students have the same architecture as their corresponding teachers, except that each student additionally contains a predictor [11], whose job is to predict the teacher targets from the other modality. The video backbone is a CSN [29] and the audio a ResNet-18 [15]; the temporal strides are modified to output 25 embeddings per second. For the predictors, we use 1-block transformers [9] to allow modelling of temporal information. The video-to-audio loss, $\mathcal{L}_{v\rightarrow a}$, is the cosine similitary between the video predictor outputs and the outputs of the audio teacher projectors. $\mathcal{L}_{a\rightarrow v}$ is defined similary. The total loss is $\mathcal{L} = \mathcal{L}_{v\rightarrow a} + \mathcal{L}_{a\rightarrow v}$. The students are optimised via gradient descent with a stop-gradient operation on the teachers, and the teachers are exponential moving averages of the students [2, 11].

Stage 2: multi-task forgery detection. Since we aim to obtain a *visual-only* forgery detector, we now discard the audio networks. We propose to use the video teacher from stage 1 to produce targets for our detector to predict. At the same time, the network performs forgery detection, in a multi-task fashion. The video student from stage 1 is used to initialise the backbone. Note that the teacher is frozen in this stage. Our framework encourages the network to classify real and fake videos by focusing on high-level spatio-temporal characteristics of facial appearance and behaviour.

We again use our dataset of real faces, but we now also assume access to a dataset of fake videos. Our architecture consists of a shared backbone and two heads: a supervised head for the forgery classification loss and an auxiliary one for the target prediction loss. The auxiliary loss, $\mathcal{L}_a$, is the cosine similarity between the predictor outputs and the video teacher representations. The supervised loss, $\mathcal{L}_s$, is a logit-adjusted version of binary cross entropy, as proposed in [23], to address any class imbalance. The final loss is $\mathcal{L} = \mathcal{L}_s + \mathcal{L}_a$.

Method	CDF	DFDC	FSh	DFo	Avg
Xception [27]	73.7	70.9	72.0	84.5	75.3
CNN-aug [30]	75.6	72.1	65.7	74.4	72.0
Patch-based [3]	69.6	65.6	57.8	81.8	68.7
Face X-ray [19]	79.5	65.5	92.8	86.8	81.2
CNN-GRU [28]	69.8	68.9	80.8	74.1	73.4
Multi-task [25]	75.7	68.1	66.0	77.7	71.9
DSP-FWA [20]	69.5	67.3	65.5	50.2	63.1
Two-branch [22]	76.7	—	—	—	—
LipForensics [14]	82.4	73.5	97.1	97.6	87.7
FTCN [32]	**86.9**	74.0	98.8	98.8	89.6
CSN	69.4	68.1	87.9	89.3	78.7
RealForensics (ours)	**86.9**	**75.9**	**99.7**	**99.3**	**90.5**

Table 1. **Cross-dataset generalisation.** AUC scores (%) on CelebDF-v2 (CDF), DeepFake Detection Challenge (DFDC), FaceShifter (FSh), and DeeperForensics (DFo), after training on FaceForensics++. Best results are in **bold**.

Method	Noise	Blur	Pixel	Compress	Avg
Xception [27]	53.8	60.2	74.2	62.1	62.6
CNN-aug [30]	54.7	76.5	91.2	72.5	73.7
Patch-based [3]	50.0	54.4	56.7	53.4	53.6
Face X-ray [19]	49.8	63.8	88.6	55.2	64.4
CNN-GRU [28]	47.9	71.5	86.5	74.5	70.1
LipForensics [14]	73.8	**96.1**	95.6	95.6	90.3
FTCN [32]	53.1	95.8	98.2	86.4	83.4
RealForensics (ours)	**79.7**	95.3	**98.4**	**97.6**	**92.8**

Table 2. **Robustness to common corruptions.** Average AUC scores (%) across five intensity levels for corruption types proposed in [16], as well as the average score across all corruptions.

3. Main results

Cross-dataset generalisation. We train on FaceForensics++ (FF++) [27] and then test on unseen datasets: CelebDF-v2 [21], DFDC [8], FaceShifter [18], and DeeperForensics [16]. The video-level AUC results are given in Table 1. Our detector obtains state-of-the-art performance without (1) using auxiliary labelled supervision [14], (2) heavily constraining the network by freezing large parts [14] or removing spatial convolutions [32], nor (3) using audio at test-time [34]. We also outperform the baseline of training a CSN [29] network on the forgery data, indicating the effectiveness of leveraging real data using our approach.

Robustness to common corruptions. We assess robustness to *unseen* perturbations. The set of perturbations, proposed in [16], are Gaussian noise and blur, pixelation, and video compression. Each perturbation type is applied at five intensity levels on raw FF++ samples. Table 2 presents the average video-level AUC across all intensity levels for each corruption type. RealForensics suffers significantly less from common corruptions than frame-based methods that target low-level cues, such as [3, 19], and also outperforms related video-based methods LipForensics and FTCN [32].

4. Conclusion

We propose RealForensics, an approach that uses large amounts of unlabelled real data to detect fake videos. We have shown that our method simultaneously achieves strong cross-manipulation generalisation and robustness to common corruptions. We hope our study encourages future research on leveraging real faces for robust forgery detection.

Acknowledgements. This work has been supported in part by Meta Platforms through research funding made available directly to Imperial College London (project P93445: Cross-modal learning of emotions). Alexandros Haliassos was financially supported by an Imperial President's PhD Scholarship. All training, testing, and ablation studies have been conducted at Imperial College.

References

[1] 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. 1

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 2

[3] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *European Conference on Computer Vision*, pages 103–120. Springer, 2020. 1, 2

[4] Zhikai Chen, Lingxi Xie, Shanmin Pang, Yong He, and Bo Zhang. Magdr: Mask-guided detection and reconstruction for defending deepfakes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9014–9023, 2021. 1

[5] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer, 2016. 2

[6] 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. 1

[7] 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. 1

[8] 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. 1, 2

[9] 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. *arXiv preprint arXiv:2010.11929*, 2020. 2

[10] Sheldon Fung, Xuequan Lu, Chao Zhang, and Chang-Tsun Li. Deepfakeucl: Deepfake detection via unsupervised contrastive learning. *arXiv preprint arXiv:2104.11507*, 2021. 1

[11] 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, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 1, 2

[12] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3473–3481, 2021. 1

[13] 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. 1

[14] 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

[15] 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. 2

[16] 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. 1, 2

[17] Sohail Ahmed Khan and Hang Dai. Video transformer for deepfake detection with incremental learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1821–1828, 2021. 1

[18] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5074–5083, 2020. 1, 2

[19] 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. 1, 2

[20] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019. 2

[21] 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. 1, 2

[22] 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. 2

[23] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020. 2

[24] 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. 1

[25] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*, 2019. 2

[26] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, pages 86–103. Springer, 2020. 1

[27] 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, 2

[28] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, 3(1), 2019. 2

[29] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019. 2

[30] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8695–8704, 2020. 1, 2

[31] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2185–2194, 2021. 1

[32] 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. 2, 3

[33] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Two-stream neural networks for tampered face detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1831–1839. IEEE, 2017. 1

[34] 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. 2

[35] Xiangyu Zhu, Hao Wang, Hongyan Fei, Zhen Lei, and Stan Z Li. Face forgery detection by 3d decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2929–2939, 2021. 1