SAFE: Sequential Attentive Face Embedding with Contrastive Learning for Deepfake Video Detection

Juho Jung
jhjeon9@g.skku.edu
Sungkyunkwan University
Seoul, South Korea

Chaewon Kang
codnjs3@g.skku.edu
Sungkyunkwan University
Seoul, South Korea

Jeewoo Yoon
yoonjeewoo@g.skku.edu
Raondata
Seoul, South Korea

Simon S. Woo
swoo@g.skku.edu
Sungkyunkwan University
Suwon, South Korea

Jinyoung Han*
jinyounghan@skku.edu
Sungkyunkwan University
Seoul, South Korea

ABSTRACT

The emergence of hyper-realistic deepfake videos has raised significant concerns regarding their potential misuse. However, prior research on deepfake detection has primarily focused on image-based approaches, with little emphasis on video. With the advancement of generation techniques enabling intricate and dynamic manipulation of entire faces as well as specific facial components in a video sequence, capturing dynamic changes in both global and local facial features becomes crucial in detecting deepfake videos. This paper proposes a novel sequential attentive face embedding, SAFE, that can capture facial dynamics in a deepfake video. The proposed SAFE can effectively integrate global and local dynamics of facial features revealed in a video sequence using contrastive learning. Through a comprehensive comparison with the state-of-the-art methods on the DFDC (Deepfake Detection Challenge) dataset and the FaceForensic++ benchmark, we show that our model achieves the highest accuracy in detecting deepfake videos on both datasets.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; • **Security and privacy** → **Software and application security**.

KEYWORDS

Deepfake Video Detection, Contrastive Learning, Sequential Embedding, Face Embedding

ACM Reference Format:
Juho Jung, Chaewon Kang, Jeewoo Yoon, Simon S. Woo, and Jinyoung Han. 2023. SAFE: Sequential Attentive Face Embedding with Contrastive Learning for Deepfake Video Detection. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3583780.3615279

*Corresponding Author.

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom.

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3615279

1 INTRODUCTION

The emergence of advanced deepfake generation techniques [25] (e.g., Face2Face [38], FaceSWAP [19], Face-Forensics [29]) has opened up new possibilities in various domains, including film and healthcare, by enabling the visualization of concepts, which have been challenging so far. However, the misuse and malicious applications of deepfake technology, such as impersonating someone in online interviews [26] or creating pornographic videos by superimposing the faces of celebrities [41], have raised significant social concerns.

Hence, there have been great efforts on deepfake detection using deep learning techniques [3, 8, 33, 36, 45]. So far, most of prior deepfake detection methods have primarily focused on image-based approaches [2, 13, 45], detecting deepfake in a given image or in a snapshot of a given video. To identify the unnaturalness shown in a deepfake image or a frame in a video, either local facial features (i.e., detailed information about specific facial regions such as facial landmarks or organs) [9, 21, 23, 32] or global facial features (i.e., overall appearance and structure of the face represented as a vector) [22, 24, 27, 42] have been utilized. However, with the advancement of video generation techniques, deepfake videos can have various and frame-by-frame manipulations, ranging from subtle changes in specific facial components to entire face swaps [21]. Here, relying solely on global or local facial features may not capture sequential dynamics across frames in a video [18].

To address this issue, we propose a novel sequential-attentive face embedding, SAFE, that can capture the temporal dynamics of video frames for detecting deepfake videos. The proposed model introduces a cross-attention mechanism for combining global and local facial features across multiple frames in a video. For highlighting the difference between fake and real videos, we apply contrastive learning [5, 39], which has been used for detecting fake images by assessing the similarity between manipulated and original images [11, 12, 31]. Specifically, we extend to leverage contrastive learning in deepfake video detection by integrating the frame-by-frame and cross-attentive global and local facial features.

The proposed model, SAFE, was evaluated on popular deepfake video datasets, DFDC (Deep Fake Detection Challenge) and FaceForensic++, and outperforms other state-of-the-art methods. An ablation study highlights that incorporating both global and local facial features achieved through leveraging contrastive learning improves performances compared to using a single feature alone. To the best of our knowledge, this is the first attempt to leverage the

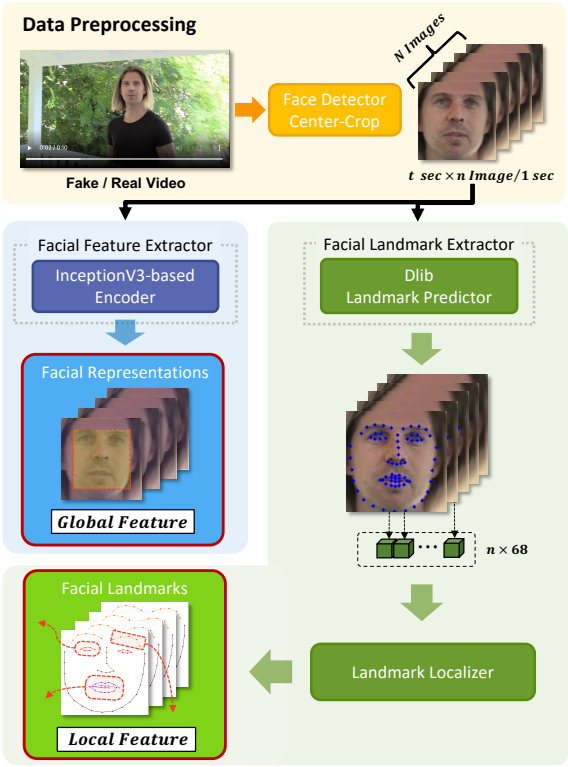

Figure 1: Facial feature extraction process for detecting deepfake videos.

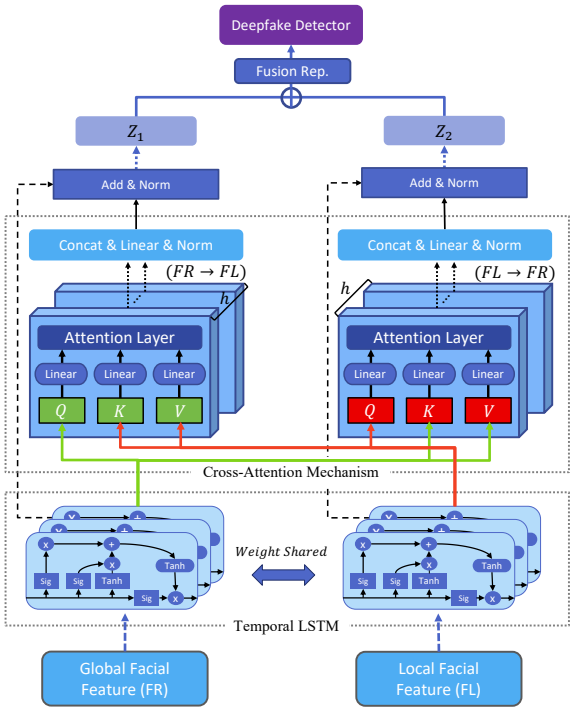

Figure 2: Overall architecture of sequential attentive face embedding. (FR = Facial Representations as a global facial feature, FL = Facial Landmarks as a local facial feature)

temporal dynamics of global and local facial features for detecting deepfake videos.

2 METHODOLOGY

2.1 Deepfake Video Dataset

Deepfake Detection Challenge Dataset (DFDC): To train and evaluate the deepfake detection models, we use the Facebook AI Deepfake Detection Challenge Dataset [10] [1]. The dataset includes 125,736 total clips obtained from 3,426 paid actors and was manipulated through diverse DeepFake and GAN-based face-swapping techniques. Due to GPU resource constraints, we randomly selected 45,801 clips from a total of 125,736 clips for training models. We used the same validation set consisting of 4,000 clips and the same test set with 5,000 clips, provided by the DFDC for evaluation.

FaceForensic++: To validate whether the proposed method is generally applicable, we use FaceForencis++ [29] (an extend version of FaceForensics [28]) in our experiments. The FaceForensic++ dataset consists of 1000 original videos from YouTube and 4000 manipulated videos created through different forgery methods (e.g., DeepFake [2], FaceSwap [3], Face2Face [38], and NeuralTexture [37]). Among them, we use the DeepFakes category, which includes 1000 manipulated videos and 1000 original videos, for our analysis.

[1]https://ai.facebook.com/datasets/dfdc
[2]https://github.com/deepfakes/faceswap
[3]https://github.com/MarekKowalski/FaceSwap/

2.2 Facial Feature Extraction

Figure 1 illustrates the overall feature extraction process. The given video clip C is first split into seconds using the OpenCV [4] video capture function. Since the duration (D) varies across video clips, we obtain N frames $\{c_1, c_2, \cdots, c_N\}$, where $N = D \times t(FPS)$. To ensure consistency, we set $t = 1$ (one frame per second = 1 FPS). We then detect a face in each frame $c \in \{c_1, c_2, \cdots, c_N\}$ and perform center-cropping to obtain a cropped face region with a size of 128 x 128 pixels. In cases where a face is not detected, we replace it with zero-vectors. As a result, we obtain a (N, 128, 128, 3) shape of cropped facial images. To extract **global facial feature**, we employ an Inception-V3-based [6] encoder. Initially, we extract 128-dimensional feature vectors for each cropped face image using the weight-sharing encoder. These feature vectors are then concatenated to form a global facial feature with dimensions $N \times 128$. For **local facial feature**, we leverage Dlib [16], a widely-used open-source software for facial analysis. This landmark predictor extracts 68 facial landmarks in each center-cropped face image, which are then fed into the landmark localizer. The landmark localizer further categorizes the 68 facial landmarks into seven distinct facial organs based on their (x, y) coordinates: jaw ([0:17]), right eyebrow ([17:22]), left eyebrow ([22:27]), nose ([27:36]), right eye ([36:42]), left eye ([42:48]), and mouth ([48:68]). By zero-padding and concatenating each x and y coordinates, our generated local facial feature exhibits a shape of $N \times 7 \times 40$.

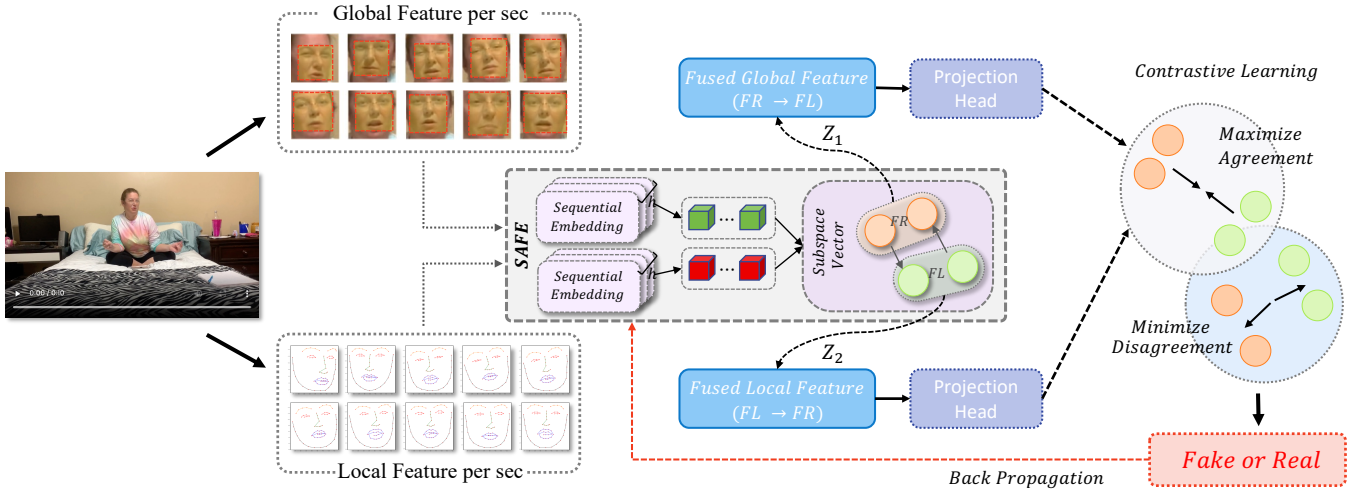

Figure 3: SAFE with contrastive learning.

2.3 Sequential Attentive Face Embedding (SAFE)

To capture global and local frame-by-frame characteristics of each video, we propose sequential attentive face embedding (SAFE), as shown in Figure 2. For representing facial representation (FR) as a global feature and facial landmarks (FL) as a local feature, we use two separate temporal LSTM (long short-term memory [14]) modules. These modules are designed to capture sequential patterns present in a given video sequence. We then introduce a cross-attention mechanism [44] with h multi-heads that can jointly consider FR and FL, capturing mutual information between them. The cross-attention mechanism is defined by Equations 1 and 2, where Q, K, V, and W represent the query, key, value, and a learnable parameters weight matrix, respectively.

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W \tag{1}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2}$$

We set $Q = FR$ and $K = V = FL$ for contextual extraction from FL and incorporate it into FR ($FR \rightarrow FL$). An opposite alignment strategy ($FL \rightarrow FR$) was applied to obtain latent information from both directions. By concatenating these two representations, the model comprehends the mutual dependencies between FR and FL, which provide a deeper comprehension of a video from both global and local perspectives. For single usage of FR or FL, cross-attention shifts to self-attention. In the final stage, the Deepfake Detector in SAFE incorporates fully connected layers, dropout layers with $p = 0.4$, and softmax activation functions to classify whether a given video is real or fake.

2.4 SAFE with Contrastive Learning

For highlighting the difference between fake and real videos, we apply contrastive learning [5, 39] in SAFE as illustrated in Figure 3. We use Z_1 and Z_2 in Figure 2, which can be obtained before the concatenation for the final fusion representation. Initially, we feed Z_1 and Z_2 through a projection head, which consists of multiple linear layers. This step enables the transformation of the input vectors into a suitable representation for contrastive learning. To measure the similarity between projected Z_1 and Z_2, we use the cosine similarity as the loss function as follows:

$$sim(Z_1, Z_2) = \frac{Z_1 \cdot Z_2}{\|Z_1\|_2 \cdot \|Z_2\|_2} \tag{3}$$

We employ the categorical cross-entropy as a loss function, which enables effective backpropagation by leveraging contrastive loss as follows:

$$Loss(z_i, z_j) = -\log \frac{exp(sim(z_i, z_j))}{\sum_{k=1}^{2} exp(sim(z_i, z_k))} \tag{4}$$

$$L = L^{task} + Loss(Z_1, Z_2) + Loss(Z_2, Z_1) \tag{5}$$

where L^{task} represents the classification task of distinguishing between fake and real videos.

3 EXPERIMENTS

3.1 Experimental Settings

The proposed model and baselines are trained using Adam [17] optimizer, with a learning rate 0.001. In addition, we set the dropout and epochs to 0.4 and 100, respectively. Note that all weights are randomly initialized in the proposed and baseline models. We evaluate the deepfake detection performance based on the following three metrics: accuracy, precision, and recall.

Baselines: For comparing the proposed models and baselines on ***DFDC***, we choose the eight popular methods that are known as accurate in deepfake detection [20, 22, 27, 40, 47]: (i) **EfficientNetB5** [35], (ii) **Ensemble of CNNs** [3], (iii) **TD-3DCNN** [46], (iv) **Xception** [6], (v) **InceptionV3** [34], (vi) **MesoInception** [1], (vii) **Conv-LSTM** [13], and (viii) **Conv-LSTM + MTCNN** [27]. For evaluation on ***FaceForensic++***, we choose the three high-performing methods on *DFDC* from Table 1 : (i) **Xception** [6], (ii) **MesoInception** [1], and (iii) **Conv-LSTM** [13]. In addition, we further evaluate four high-performing methods [29, 45] on *FaceForensic++*: (i) **CNN + GRU + STN** [30], (ii) **Face X-ray** [22], (iii) **FaceCatcher** [7], and (iv) **Local CNN** [36].

Table 1: Overall performance of the proposed model and the baselines on DFDC. (FR = Facial Representations, FL = Face Landmarks, CL = Contrastive Learning)

Method	Input	Accuracy	Precision	Recall
EfficientNetB5 [35]	FR	0.842	0.838	0.657
	FL	0.849	**0.898**	0.825
	Both	**0.876**	0.819	**0.865**
Ensemble of CNNs [3]	FR	0.869	0.867	0.880
	FL	0.836	0.783	0.865
	Both	**0.881**	**0.900**	**0.895**
TD-3DCNN [46]	FR	0.872	0.869	0.882
	FL	0.828	0.743	**0.906**
	Both	**0.885**	**0.905**	0.878
Xception [29]	FR	0.872	0.889	0.868
	FL	0.840	**0.904**	0.797
	Both	**0.913**	0.901	**0.910**
InceptionV3 [34]	FR	0.873	0.901	0.860
	FL	0.864	0.850	0.883
	Both	**0.905**	**0.910**	**0.904**
MesoInception [1]	FR	0.884	0.874	0.899
	FL	0.867	0.876	0.869
	Both	**0.915**	**0.913**	**0.904**
Conv-LSTM [13]	FR	0.876	**0.905**	0.862
	FL	0.848	0.898	0.825
	Both	**0.886**	**0.905**	**0.878**
Conv-LSTM + MTCNN [27]	FR	0.888	0.908	0.879
	FL	0.882	0.876	**0.894**
	Both	**0.906**	**0.931**	0.892
SAFE	FR	0.898	0.908	0.879
	FL	0.882	0.876	0.940
	Both	**0.931**	**0.911**	**0.921**
Conv-LSTM [13] with CL	Both	0.916	0.921	0.897
MesoInception [1] with CL	Both	0.929	0.934	0.913
SAFE with CL	Both	**0.957**	**0.954**	**0.927**

3.2 Experimental Results

3.2.1 Overall Performance. We evaluate the performance of the proposed model on **DFDC**, comparing it with the eight baseline methods. As shown in Table 1, our proposed method, **SAFE**, outperforms other baselines. The result indicates that the proposed cross-attention mechanism is effective in integrating local facial information and global high-level information in detecting deepfake videos. Among the baseline models, **Xception** [6] and **MesoInception** [1] show high performance; MesoInception prioritizes the extraction of mesoscopic image features [15], which lies in between high- and low-level features. This indicates that mesoscopic features can be useful in detecting deepfake videos [43].

3.2.2 Analysis on Feature Importance. To assess the significance of each feature (i.e., face representations and landmarks) in detecting deepfake videos, we conduct a performance analysis on models trained with each feature. As shown in Table 1, most models, including the baselines and SAFE, trained only with facial representations, exhibit a higher performance than those trained only with face landmarks. This suggests that high-dimensional latent representations are more useful than detailed morphological information in deepfake detection. Furthermore, we find a notable improvement when both features are well-combined through a cross-attention mechanism. This implies that learning both facial representations and face landmarks, along with their inter-dependencies, is more effective than relying solely on a single feature.

Table 2: Deepfake video detection performance on Faceforensic++ (Deepfake) dataset.

Method	Accuracy
CNN + GRU + STN [30]	0.969
Face X-ray [22]	0.989
FaceCatcher [7]	0.938
Local CNN [36]	0.979
Xception [29]	0.962
MesoInception [1]	0.984
Conv-LSTM [13]	0.953
SAFE	**0.992**
SAFE with CL	**0.995**

3.2.3 Ablation Study on Contrastive Learning. We show that applying contrastive learning to SAFE shows higher performance than SAFE w/o contrastive learning, as shown in Table 1. We also find that high-performing baselines, Conv-LSTM and MesoInception, also show an enhanced performance by incorporating contrastive learning. This highlights the positive impact of contrastive learning on enhancing the comprehension and integration of multiple representations in deepfake video detection.

3.2.4 Evaluation on FaceForensic++. To validate the generalizability of our model to another deepfake video dataset, we conduct an evaluation using a popular deepfake dataset, Faceforensic++. As shown in Table 2, our proposed model, SAFE, outperforms other state-of-the-art models and achieves an high accuracy of 0.995 on Faceforensic++. This demonstrates the general use of the proposed model in detecting deepfake videos.

4 CONCLUSION

In this paper, we proposed SAFE, a novel deepfake video detection model that effectively captures temporal dynamics. SAFE employs a cross-attention mechanism to combine global and local facial features, enabling differentiation between fake and real videos while utilizing contrastive learning to emphasize their distinctions. Experimental results on DFDC and FaceForensic++ demonstrated the superior performance of SAFE compared to the state-of-the-art methods. Furthermore, our analysis emphasizes the significant contribution of contrastive learning in enhancing deepfake video detection. Our work has important implications for identifying, addressing, and preventing potential social issues arising from the proliferation of deepfake videos.

ACKNOWLEDGMENTS

This research was supported by the National Research Foundation (NRF) of Korea grant funded by the Korea government (MSIT) (No. 2021R1A4A3022102, No. 2023R1A2C2007625), and Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2023-00230337, Advanced and Proactive AI Platform Research and Development Against Malicious Deepfakes).

REFERENCES

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 1–7.

[2] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. 2019. Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*. 0–0.

[3] Nicolo Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. 2021. Video face manipulation detection through ensemble of cnns. In *2020 25th international conference on pattern recognition (ICPR)*. IEEE, 5012–5019.

[4] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[6] François Chollet. 2017. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1251–1258.

[7] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. 2020. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence* (2020).

[8] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. 2022. Combining efficientnet and vision transformers for video deepfake detection. In *Image Analysis and Processing–ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part III*. Springer, 219–229.

[9] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. 2021. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15108–15117.

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

[11] Fengkai Dong, Xiaoqiang Zou, Jiahui Wang, and Xiyao Liu. 2023. Contrastive learning-based general Deepfake detection with multi-scale RGB frequency clues. *Journal of King Saud University-Computer and Information Sciences* 35, 4 (2023), 90–99.

[12] Sheldon Fung, Xuequan Lu, Chao Zhang, and Chang-Tsun Li. 2021. Deepfakeucl: Deepfake detection via unsupervised contrastive learning. In *2021 international joint conference on neural networks (IJCNN)*. IEEE, 1–8.

[13] David Güera and Edward J Delp. 2018. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*. IEEE, 1–6.

[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[15] Loc Huynh, Weikai Chen, Shunsuke Saito, Jun Xing, Koki Nagano, Andrew Jones, Paul Debevec, and Hao Li. 2018. Mesoscopic facial geometry inference using deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8407–8416.

[16] Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.

[17] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[18] Donggeun Ko, Sangjun Lee, Jinyong Park, Saebyeol Shin, Donghee Hong, and Simon S Woo. 2022. Deepfake Detection for Facial Images with Facemasks. *arXiv preprint arXiv:2202.11359* (2022).

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

[20] Sangjun Lee, Donggeun Ko, Jinyong Park, Saebyeol Shin, Donghee Hong, and Simon S Woo. 2022. Deepfake Detection for Fake Images with Facemasks. In *Proceedings of the 1st Workshop on Security Implications of Deepfakes and Cheapfakes*. 27–30.

[21] John K Lewis, Imad Eddine Toubal, Helen Chen, Vishal Sandesera, Michael Lomnitz, Zigfried Hampel-Arias, Calyam Prasad, and Kannappan Palaniappan. 2020. Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*. IEEE, 1–9.

[22] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 2020. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5001–5010.

[23] Meng Li, Beibei Liu, Yongjian Hu, Liepiao Zhang, and Shiqi Wang. 2021. Deepfake detection using robust spatial and temporal features from facial landmarks. In *2021 IEEE International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 1–6.

[24] Xurong Li, Kun Yu, Shouling Ji, Yan Wang, Chunming Wu, and Hui Xue. 2020. Fighting against deepfake: Patch&pair convolutional neural networks (PPCNN). In *Companion Proceedings of the Web Conference 2020*. 88–89.

[25] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. 2023. Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence* 53, 4 (2023), 3974–4026.

[26] Vineet Mehta, Parul Gupta, Ramanathan Subramanian, and Abhinav Dhall. 2021. Fakebuster: a deepfakes detection tool for video conferencing scenarios. In *26th International Conference on Intelligent User Interfaces-Companion*. 61–63.

[27] Daniel Mas Montserrat, Hanxiang Hao, Sri K Yarlagadda, Sriram Baireddy, Ruiting Shao, János Horváth, Emily Bartusiak, Justin Yang, David Guera, Fengqing Zhu, et al. 2020. Deepfakes detection with automatic face weighting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 668–669.

[28] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2018. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179* (2018).

[29] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1–11.

[30] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. 2019. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)* 3, 1 (2019), 80–87.

[31] Dongyao Shen, Youjian Zhao, and Chengbin Quan. 2022. Identity-Referenced Deepfake Detection with Contrastive Learning. In *Proceedings of the 2022 ACM Workshop on Information Hiding and Multimedia Security*. 27–32.

[32] Pu Sun, Yuezun Li, Honggang Qi, and Siwei Lyu. 2020. Landmark breaker: obstructing deepfake by disturbing landmark extraction. In *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 1–6.

[33] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang. 2014. Deep learning face representation by joint identification-verification. *Advances in neural information processing systems* 27 (2014).

[34] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.

[35] Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*. PMLR, 6105–6114.

[36] Michail Tarasiou and Stefanos Zafeiriou. 2020. Extracting deep local features to detect manipulated images of human faces. In *2020 IEEE international conference on image processing (ICIP)*. IEEE, 1821–1825.

[37] Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)* 38, 4 (2019), 1–12.

[38] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2387–2395.

[39] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? *Advances in neural information processing systems* 33 (2020), 6827–6839.

[40] Ruben Tolosana, Sergio Romero-Tapiador, Julian Fierrez, and Ruben Vera-Rodriguez. 2021. Deepfakes evolution: Analysis of facial regions and fake detection performance. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part V*. Springer, 442–456.

[41] MC Weerawardana and TGI Fernando. 2021. Deepfakes detection methods: A literature survey. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*. IEEE, 76–81.

[42] Deressa Wodajo and Solomon Atnafu. 2021. Deepfake video detection using convolutional vision transformer. *arXiv preprint arXiv:2102.11126* (2021).

[43] Zhiming Xia, Tong Qiao, Ming Xu, Xiaoshuai Wu, Li Han, and Yunzhi Chen. 2022. Deepfake Video Detection Based on MesoNet with Preprocessing Module. *Symmetry* 14, 5 (2022), 939.

[44] Jeewoo Yoon, Chaewon Kang, Seungbae Kim, and Jinyoung Han. 2022. D-vlog: Multimodal Vlog Dataset for Depression Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12226–12234.

[45] Peipeng Yu, Zhihua Xia, Jianwei Fei, and Yujiang Lu. 2021. A survey on deepfake video detection. *Iet Biometrics* 10, 6 (2021), 607–624.

[46] Daichi Zhang, Chenyu Li, Fanzhao Lin, Dan Zeng, and Shiming Ge. 2021. Detecting Deepfake Videos with Temporal Dropout 3DCNN.. In *IJCAI*. 1288–1294.

[47] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. 2021. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*. 15023–15033.