Improving Deepfake Video Detection with Comprehensive Self-consistency Learning

Heng Bao[1], Lirui Deng[2], Jiazhi Guan[2], Liang Zhang[3(✉)], and Xunxun Chen[3]

[1] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
baoheng@iie.ac.cn
[2] Department of Computer Science and Technology, Tsinghua University, Beijing, China
{dlr18,guanjz20}@mails.tsinghua.edu.cn
[3] CNCERT/CC, Beijing, China
zl@isc.org.cn

Abstract. Deepfake videos created by generative-base models have become a serious societal problem recently as been hardly distinguishable by human eyes, which has aroused a lot of academic attention. Previous researches have made effort to address this problem by various schemes to extract visual artifacts of non-pristine frames or discrepancy between real and fake videos, where the patch-based approaches are shown to be promising but mostly used in frame-level prediction. In this paper, we propose a method that leverages comprehensive consistency learning in both spatial and temporal relation with patch-based feature extraction. Extensive experiments on multiple datasets demonstrate the effectiveness and robustness of our approach by combines all consistency cue together.

Keywords: Deepfake detection · Digital forensics · Video classification

1 Introduction

"Seeing is believing" is hardly true in present days with the prosperity of computer science and information technology, especially the massively emerging applications of artificial intelligence. Although image and video forgery is never a new topic since the beginning of photography, open source applications represented by Deepfakes [7] and others have brought this problem into a whole new level. Face manipulation in visual content has become a effortless task with the help of deep learning based generative models like variational autoencoders (VAEs) [16] and generative adversarial networks (GANs) [12], that anyone can produce fake videos with false identity or manipulated expressions and movements (known as "Deepfake Videos") in several minutes without expert knowledge. Some of them have already been found to create malicious videos that

W. Lu et al. (Eds.): CNCERT 2022, CCIS 1699, pp. 151–161, 2022.
https://doi.org/10.1007/978-981-19-8285-9_11

violate citizen privacy or attack public figures like fake pornography and blackmailing, which may easily lead to catastrophic results with today's mass media and social networks. The detection of deepfake videos has become an hot and urgent issue.

Various methods have been proposed in recent years by academic community to effectively recognize this particular type of forged images and videos. Since the majority forgery methods share a common image stitching pipeline including face detection, warping and blending, early researches address this task by detecting suspicious artifacts left in the stitching process within frame-level, such as face warping artifacts [20] and blending boundaries [18]. To yield a result for whole video clip from frame-level prediction, they usually cascade frame-level model with a merge module, or sometimes simply use weighted average. But ignoring the dependency among consecutive frames tends to produce sub-optimal combination. Frequency-based approached [9,25] have also been included to fully utilized temporal relation. Self-consistency is another crucial concept in image forensic [14,35], where patch-based and feature-map based method have all shown promising results. Although the detection accuracy on datasets has improved significantly with different approaches presented, forgery techniques are also evolving on reducing these artifacts, which forms an ever-changing arms race.

In this work, we aim to catch both the intra-frame discrepancy during image stitched and the inherent flaws of inter-frame disalignment for more effective and robust deepfake detection. Our contributions can be summarized as follows:

- We propose a comprehensive self-consistency learning(CSCL) model to explore the intrinsic discernible evidence between pristine and deepfake videos with both spatial and temporal consistency learning.
- To achieve more effective and robust deepfake detection, we also proposed C^3Loss, namely comprehensive consistency coordination loss, which tackles the inherent defects within deepfake producing pipeline as been created frame-by-frame without sequential knowledge.
- Experiments conducted on multiple datasets demonstrate the effectiveness and robustness of our approach. Especially, best performance is reported in cross-dataset and low quality tests.

2 Related Work

Frame-Level Detection Methods. The emerging of deepfake videos on the internet raise a lot of concern to both industry and government in the past few years. Early researches [1,26] tend to address this problem by a simple classification model with a well-designed backbone. And some [20] simulated the generation process of deep forgery to better obtain artifact of fake video pipeline. Not only in academic society, a one-million bounty real-world deepfake detection competition was held by Facebook with the concern of its endangerment of social media to encourage optimal deepfake detection methods being proposed. Plenty of classification model was proposed and achieve really amazing results

beyond expectation. The winner of this competition [27] adopts the state-of-the-art image classification backbone efficientNet [28] as the main component of his model, and a novel data argumentation strategy contributes a lot to his final ranking. The runner-up of this competition fulfilled their method afterward [34], which treat the deepfake detection task as a fine-grained classification task and explicitly refine attention maps by regional independence loss. Merging with texture features extracted at the front-end layer, their model achieve state-of-the art results in several datasets. Attention map prediction scheme is also considered by [6]. In their work, forged area is predicted in both learning-base and dictionary learning ways, binary classification and attention map regression tasks are trained using a multi-task loss function.

The above mentioned detection methods are all concentrate on the RGB-domain of deepfakes, and there are some other works try to explore fake clues inherent in the frequency domain of deepfake images. Discrete cosine transform (DCT) [2] is adopted in [25], frequency layout of image is fully handled in both global and local views. In combine with learnable frequency-aware component, nonaligned infomation can be reliably detected at frame-level. Frank *et al.* [9] also leverage DCT in detection, and analysis which part makes synthetical deepfake image detectable. Their results suggest that up-sampling blocks left unique fingerprint, but those frequency clues are not robust to perturbation.

Besides, some other approaches tried to inspect artifacts from the side-view. FakeSpotter [31] do not directly use the features extracted by backbone network, but regard the neuron behaviors as the basis of discrimination, which is aimed to achieve more robust detection. To better leveraging the time factor into consideration in video-level authentication, spotting bio-metrics clues like eye blinking [19,32] and head posing sequential [32] is the first and most natural insight. DeepRhythm [24] exposes deepfake counterfeits by monitoring the heartbeat rhythms associated with minuscule periodic changes of skin color due to blood pumping through the face.

Video-level Detection Methods. Most video-level methods regard video as set of independent frames, and simply take the average confidence score of frames as the basis of judging the authenticity of video. Those methods actually follow the frame-level perspective, and neglect the interconnection between successive frames.

Güera *et al.* [13] adopt a natural way to leverage both advantages of convolutional neural network (CNN) and recurrent neural network (RNN) by using CNN for per-frame feature extraction, and RNN for temporal inconsistencies exploration. But inter-frame inconsistency modeling is not well considered in their approach. Tariq *et al.* [29] consider the artifacts introduced by the non-consecutive frames, and developed a convolutional LSTM-base residual network to achieve temporal feature learning. Basic features of the human body like eye blinking and head pose moving are utilized in [19] and [32] to distinguish the real from fake. In [23], the authors leverage the relationship between visual and audio patterns extracted from the same video to determine whether it has been modified.

Video-level detection gathers more information and, in general, should deliver better performance. But strangely, video-level evaluation results in terms of ACC and AUC are somehow lower than those at the frame-level. Zi *et al.* [37] propose two models by stacking ADD block. In their experiments, ADDNet-3D report much lower detection accuracy than ADDNet-2D, about 10 percents gap at a challenging dataset. Ganiyusufoglu *et al.* [11] adopt the state-of-the art structures used in action recognition task, and evaluate their performance in deepfake detection.

Benchmark Datasets. Several comprehensive deepfake datasets were published in recent years which greatly promotes the performance of deepfake detection methods. One of the most popular dataset is FaceForensics++ (FF++) [26]. It contains two graphic based approaches, namely Face2Face [30] and Faceswap [8], and two learning based methods include Deepfakes and Natural Textures [15]. Both face swap and face reenactment are covered. Celeb-DF [21] is one of the most challenging dataset in deepfake detection task with clear identity label and pixel level annotation. During the deepfake generation stage, they scrutinizes carefully about several problems during fake video generation, including color mismatch, inaccurate face masks and video temporal flickering. With more attention drawn to this research topic, some new and better annotated datasets been proposed with more specific purpose recently like WildDeepfake [37] for real-world challenge and OpenForeinsic [17] for multiple face scenario.

3 Approach

Given an input video with certain human activity, our goal is to detect if the identity is replaced or facial expression of character is manipulated. We propose CSCL network as shown in Fig. 1 to improve the robustness and generalization ability of deepfake-style forgery video detector with the help of self-consistency by measuring the comprehensive spatial and temporal discrepancy within the image stream.

To be more precise, our method mainly exploit a comprehensive consistency which tackles the substantial drawback of deepfake videos producing pipeline:

- **Intra-frame: Spatial consistency**. Intra-frame consistency in deepfake video are mostly provided by blending algorithm like Gaussian blur or Poisson fusion, which has been proved to be distinguishable.
- **Inter-frame: Temporal consistency**. Common generative models with frame-by-frame swapping process can not guarantee a smooth temporal momentum, while most of the former consistency learning methods only focus on single manipulated frame and tend to be overfit in one subset of manipulation.
- **Comprehensive consistency coordination**. Extra blur and filtering can conceal the intra-frame discrepancy, and inter-frame consistency may also be diminished by adaptive average blending. Unlike previous work, we utilize inter- and intra-frame consistency coordination for more robust deepfake video detection.

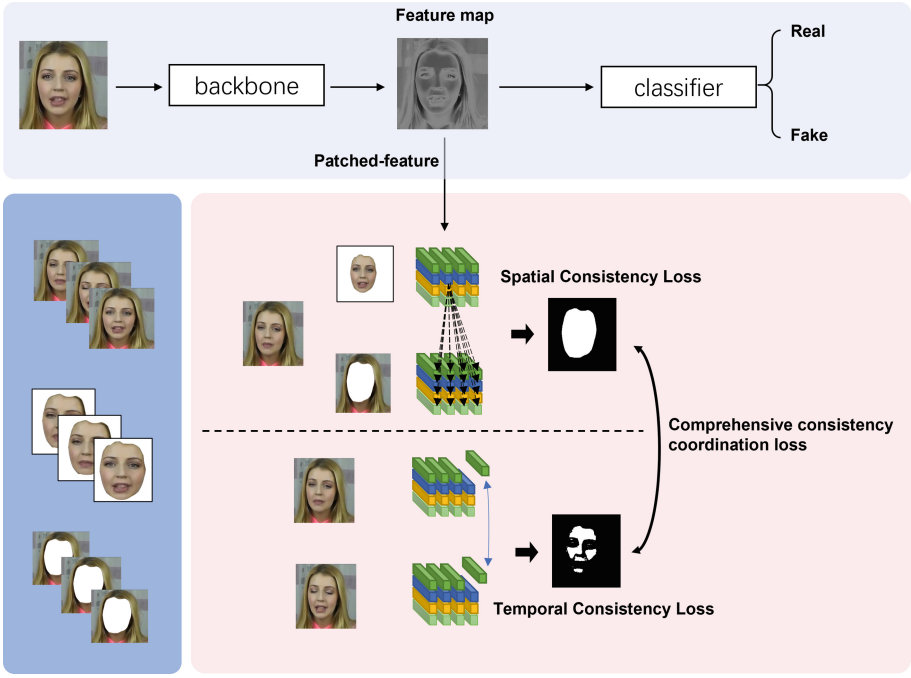

Fig. 1. Framework of CSCL network

3.1 Problem Formulation

We formulate the video-level deepfake detection task at beginning. Dataset $D = \{X_i, L_i\}_{n=1}^{N}$ consists of n pairs of video-clip and it's label with `fake` or `real` denoted as $L_i = \{0, 1\}$. Video clip can be seen as multiple consecutive frames $X_v = \{x_t\}_{t=1}^{T_v}$, where $x_t \in \mathbb{R}^{C \times H \times W}$ is the t-th frame of video X_v, and the total number of frames is denoted as T_v. All the frames in one specific video X_v are deemed as manipulated if X_v is labeled with `fake`, vice versa. The goal of deepfake detection is to learn a model Φ, which takes all consecutive frames of one video, and give a clear judgment of the authenticity, formulated as $\Phi(X_v) \in \{\texttt{fake}, \texttt{real}\}$.

3.2 Design of Model

Spatial Consistency of Contexts vs. Faces. Computing similarity scores among images patched for inconsistency has already been proved effective in image forensic researches [33,35,36]. Without loss of generality, we first obtain feature f_t of image x_t from backbone model $\mathbf{G}$ of size $H' \times W' \times C'$ where H' and W' and patch numbers along columns and rows.

$$f_t = \mathbf{G}(x_t)_{t=1}^{T_v} \in \mathbb{R}^{C' \times H' \times W'} \tag{1}$$

For each frame x_t we follow the [35] to calculate the 4D consistency map $\hat{SM}$ with:

$$\begin{aligned} \hat{SM}_{h,w,h',w'} &= d(f_t^{h,w}, f_t^{h',w'}) \\ &= 1 - \cos(f_t^{h,w}, f_t^{h',w'}) \end{aligned} \tag{2}$$

While each frame's mask have only two possible status : manipulated or not, for patch P located in face area denoted as P_f, else in context as P_c, and $\psi(P_f) = 1$ else $\psi(P_c) = 0$, the ground truth:

$$SM_{P_i,P_j} = \psi(P_i) \oplus \psi(P_j) \tag{3}$$

and the spatial consistency loss:

$$L_{SC} = |SM - \hat{SM}| \tag{4}$$

Temporal Consistency of Consecutive Frames. In order to catch inconsistency between successive frames, we further extend the attention to temporal consistency learning. As we have obtained the patch-base feature f_t from x_t, we consider the relation between f_t and f_{t-1}. For each path $P_{h,w}$ at timestamp t, we have a 2D consistency map:

$$\begin{aligned} \hat{TM}_{P_t^{h,w} P_{t-1}^{h,w}} &= d(f_t^{h,w}, f_{t-1}^{h,w}) \\ &= 1 - \cos(f_t^{h,w}, f_{t-1}^{h,w}) \end{aligned} \tag{5}$$

considering the momentum between t and $t-1$, we calculate temporal consistency loss:

$$L_{TC} = \sum_t |\hat{TM}_t - \frac{\sum_{h,w} \hat{TM}_{t,h,w}}{HW}| \tag{6}$$

Coordinating Temporal and Spatial Consistency. It's not hard to imaging that no matter in pristine or deepfake video people's face will be moving most of the time, either talking or acting expressions. Otherwise the there's no need to forge this static video which conveys no more information than just a photo. Only measuring the discrepancy of distance between consecutive face and context would yield lots of false alarm. Therefore we propose a comprehensive consistency coordination loss for adaptive learning by monitoring the relation between temporal and spatial consistency. Now we have final Loss function:

$$L = L_{real/fake} + \lambda L_{SC} + \beta L_{TC} + (1 - \beta) L_{CCC} \tag{7}$$

4 Experiment Results

Implementation Details. We modify Xception [4] as the backbones and their parameters are initialized by Xception pre-trained on ImageNet. We train our model using Adam optimizer with initial learning rate 1e-4 and weight decay 1e-7. Train epoch size is set to 2000, batch size is set to 32, and if validation loss is not getting better in 5 epochs, learning rate is decayed by factor 0.3, so that model can converge after several learning rate decays.

4.1 In-Dataset Evaluation on FF++

FF++ is one of the most popular dataset for evaluating deepfake detection methods. It contains 1000 real videos collected from internet, and 4000 fake videos generated by four kinds of deepfake techniques. More over, FF++ provides 3 different qualities of videos, we use the high quality (c23) and low quality (c40) versions in this section. The raw quality videos are not considered because they are not very common on the internet. We use the same split as [26], both real and fake video is split into train, validation and test set according to the ratio of 72:14:14. But it is noticed that number of real videos is much smaller than fake videos. So, we oversample real videos to balance the classes when training. At test stage, one video could contain several clips in FF++, we extract as much clips as we can from one video (interval is set to 16, no overlap), and take the average score of all clips as confidence score of the video. The test results are listed in Table 1.

Table 1. In-dataset Performance (ACC %) on four types of deepfake in FF++. DF: DeepFakes, F2F: Face2Face, FS: FaceSwap, NT: NeuralTextures. The best result is shown in bold text, and the second-best is underlined.

	Methods	DF	F2F	FS	NT
Frame Level	LD-CNN [10]	75.00	56.00	51.00	62.00
	Constrained Conv [5]	87.00	82.00	74.00	74.00
	CustomPooling CNN [3]	80.00	62.00	59.00	59.00
	MesoNet [1]	<u>90.00</u>	<u>83.00</u>	<u>83.00</u>	**83.00**
	Xception [4]	**96.01**	**93.29**	**96.71**	<u>79.14</u>
Video Level	PCL [35]	96.87	94.93	<u>98.44</u>	**99.58**
	PD [33]	<u>97.53</u>	<u>96.57</u>	95.01	92.55
	ours	**100.00**	**99.84**	**99.21**	<u>99.37</u>

4.2 Cross-Dataset Evaluation on Celeb-DF

The poor Generalization ability of deepfake detection is still a thorny problem, even the state-of-the-art methods suffer from drastically performance degradation when test on deepfakes generated by unseen techniques. Our method tries to formulate deepfake detection from a discrepancy discovering aspect, and achieves the best cross-dataset performance, as the results listed in Table 2. The test model is trained on FF++ low quality, follow the setting of [22] for fair comparison. It is noticed that many methods report around 100% AUC on train set, but fail to transfer to the different dataset. Our model achieve the best cross-dataset test performance, while keep the best test result on train set.

Table 2. Cross-dataset Performance (AUC %) on Celeb-DF. The best result is shown in bold text, and the second-best is underlined.

Methods	FF++	Celeb-DF
MesoNet-Inception [1]	83.00	53.60
FWA [20]	80.10	56.90
Xception-raw [4]	99.70	48.20
Xception-c23 [4]	99.70	65.30
Xception-c40 [4]	95.50	65.50
DSP-FWA [20]	93.00	64.60
Two-Branch [22]	93.18	73.41
PCL [35]	99.79	72.44
Patch-Diffusion [33]	**99.85**	<u>74.27</u>
ours	**99.85**	**77.73**

4.3 Ablation Study

This section analyzes the effectiveness of our proposed CSCL module. CSCL consist of three parts in total: the spatial consistency, temporal consistency and comprehensive consistency coordination. To further validate whether each part of comprehensive consistency can improve the generalizability, we conduct an ablation study by comparing our methods with the following variant. (1)Xception [4]: the baseline approach without using any consistency cue. (2)Xception w/ sc: we follow the setting of [35] with only spatial patch consistency loss. (3)Xception w/ tc: we use only temporal consistency loss upon baseling. (4)Ours full CSCL model with both spatial and temporal consistency, plus consistency coordination loss. Results are listed in Table 3.

Table 3. Ablation Performance in FF++. The best result is shown in bold text.

	Methods	AUC(HQ)	AUC(LQ)
Base Line	PD [33]	99.85	94.43
	PCL [35]	99.79	96.38
Ablation	Xception+SC	98.46	98.01
	Xception+TC	95.13	94.93
	Xception+SC+TC	99.46	98.16
	CSCL(SC+TC+CCC)	**99.85**	**98.21**

5 Summary

In this paper, we try to address the problem, deepfake detection, from the view of comprehensive self-consistency learning. More specifically, we propose a CSCL

model with spatial-temporal consistency learning to explicitly formulate the inherent flaws of intra- and inter-frame disalignment in deepfakes. To achieve more effective and robust deepfake detection, we also proposed C^3Loss, namely comprehensive consistency coordination loss, which tackles the inevitable artifact within deepfake producing pipeline. Extensive experiments demonstrate the superior performance of our method in deepfake detection, especially in more realistic tests like cross-dataset and low quality setting.

References

1. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: MesoNet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7. IEEE (2018)
2. Ahmed, N., Natarajan, T., Rao, K.: Discrete cosine transform. IEEE Trans. Comput. **23**(01), 90–93 (1974)
3. Bayar, B., Stamm, M.C.: A deep learning approach to universal image manipulation detection using a new convolutional layer. In: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, pp. 5–10 (2016)
4. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
5. Cozzolino, D., Poggi, G., Verdoliva, L.: Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security, pp. 159–164 (2017)
6. Dang, H., Liu, F., Stehouwer, J., Liu, X., Jain, A.K.: On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5781–5790 (2020)
7. Deepfake github. https://github.com/deepfakes/faceswap
8. Faceswap github. https://github.com/MarekKowalski/FaceSwap/
9. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: International Conference on Machine Learning, pp. 3247–3258. PMLR (2020)
10. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital images. IEEE Trans. Inf. Forensics Secur. (TIFS) **7**, 868–882 (2012)
11. Ganiyusufoglu, I., Ngô, L.M., Savov, N., Karaoglu, S., Gevers, T.: Spatio-temporal features for generalized detection of deepfake videos. arXiv preprint arXiv:2010.11844 (2020)
12. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in neural information processing systems, vol. 27 (2014)
13. Güera, D., Delp, E.J.: Deepfake video detection using recurrent neural networks. In: 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), pp. 1–6. IEEE (2018)
14. Han, X., Morariu, V., Larry Davis, P.I., et al.: Two-stream neural networks for tampered face detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 19–27 (2017)
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)

16. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
17. Le, T.N., Nguyen, H.H., Yamagishi, J., Echizen, I.: Openforensics: large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10117–10127 (2021)
18. Li, L., et al.: Face x-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5001–5010 (2020)
19. Li, Y., Chang, M.C., Lyu, S.: In Ictu oculi: exposing AI created fake videos by detecting eye blinking. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7. IEEE (2018)
20. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656 (2018)
21. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-DF: a large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3207–3216 (2020)
22. Masi, I., Killekar, A., Mascarenhas, R.M., Gurudatt, S.P., AbdAlmageed, W.: Two-branch recurrent network for isolating deepfakes in videos. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12352, pp. 667–684. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58571-6_39
23. Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., Manocha, D.: Emotions don't lie: a deepfake detection method using audio-visual affective cues. arXiv preprint arXiv:2003.06711 (2020)
24. Qi, H., et al.: Deeprhythm: exposing deepfakes with attentional visual heartbeat rhythms. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 4318–4327 (2020)
25. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: face forgery detection by mining frequency-aware clues. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12357, pp. 86–103. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_6
26. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1–11 (2019)
27. Selim, S.: Deepfake detection (DFDC) solution by selim seferbekov. https://github.com/selimsef/dfdc_deepfake_challenge
28. Tan, M., Le, Q.: EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
29. Tariq, S., Lee, S., Woo, S.S.: A convolutional LSTM based residual network for deepfake video detection. arXiv preprint arXiv:2009.07480 (2020)
30. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2Face: real-time face capture and reenactment of RGB videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2387–2395 (2016)
31. Wang, R., et al.: Fakespotter: a simple yet robust baseline for spotting AI-synthesized fake faces. arXiv preprint arXiv:1909.06122 (2019)
32. Yang, X., Li, Y., Lyu, S.: Exposing deep fakes using inconsistent head poses. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8261–8265. IEEE (2019)

33. Zhang, B., Li, S., Feng, G., Qian, Z., Zhang, X.: Patch Diffusion: a general module for face manipulation detection. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence (2022)
34. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., Yu, N.: Multi-attentional deepfake detection. arXiv preprint arXiv:2103.02406 (2021)
35. Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., Xia, W.: Learning self-consistency for deepfake detection. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 15003–15013 (2021). https://doi.org/10.1109/ICCV48922.2021.01475
36. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Neural networks for tampered face detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1831–1839. IEEE (2017)
37. Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.G.: WildDeepFake: a challenging real-world dataset for deepfake detection. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2382–2390 (2020)