Capturing the Lighting Inconsistency for Deepfake Detection

Wenxuan Wu[1], Wenbo Zhou[1(✉)], Weiming Zhang[1(✉)], Han Fang[2], and Nenghai Yu[1]

[1] University of Science and Technology of China, Hefei 230000, China
{welbeckz,zhangwm}@ustc.edu.cn
[2] National University of Singapore, 21 Lower Kent Ridge Road, Singapore, Singapore

Abstract. The rapid development and widely spread of deepfake techniques have raised severe societal concerns. Thus detecting such forgery contents has become a hot research topic. Many deepfake detection methods have been proposed in an artifacts-driven manner. They are well-designed to capture subtle artifacts of the face region in different domains. But since the lighting information is usually ignored during the forgery process, which may cause inconsistent lighting between the original face and forged one, we believe that this kind of semantic information can be useful to promote detection accuracy. In this paper, we propose a lighting inconsistency based deepfake detection method. We apply the color constancy technique to each sample and obtain a pre-processed image. Then the unique lighting information of each sample can be obtained by calculating the difference between the processed image and the original one. The lighting information will be used as an assistant channel for better detection accuracy. Extensive experiments show that our method can achieve obvious enhancements compared to the baseline method.

Keywords: Deepfake detection · Light inconsistency · Color constancy

1 Introduction

Benefiting from the development of extraordinary deep generative models, such as Generative Adversarial Networks (GAN) [1] and Variational Autoencoders (VAE) [2], deepfake technique springs up and become an arousing research topic. The ultra-realistic face images or videos are indistinguishable for human eyes, they can easily be abused for malicious purposes and eventually lead to adverse social, political and economic consequences, such as fraud, defamation and fake news [3–7].

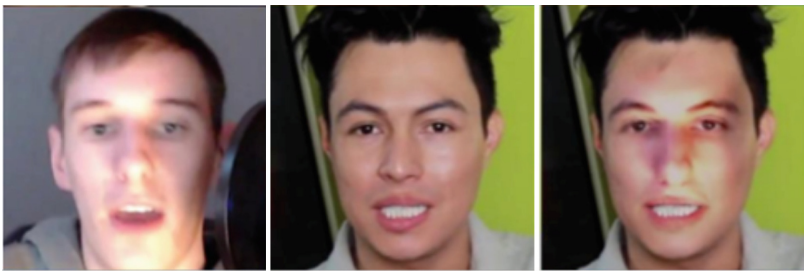
Early forged faces generated by imperfect synthesis algorithms will introduce obvious visual artifacts, which can be easily distinguished by many handcraft artifact-based detection methods [9, 10]. Some methods take the original RGB images as input and automatically train the detection models in a vanilla binary classification manner [16, 17]. In this manner, the models are likely to overfit to specific dataset and struggle to handle more challenging cases. Some recent work has carefully designed the detection model and introduced more effective modules, such as attention mechanism [18–20],

X. Sun et al. (Eds.): ICAIS 2022, LNCS 13339, pp. 637–647, 2022.
https://doi.org/10.1007/978-3-031-06788-4_52

texture features [20, 21], audio-visual modes [22] and spectrum [23], to capture more robust features for deepfake detection. However, those methods are mostly designed in an artifacts-driven way, ignoring some potential semantic inconsistency between the original faces and the forged ones.

We observe that lighting clues have not been taken into consideration in most of the deepfake forgery techniques. As shown in Fig. 1, although the forged faces can swap the identity of the source image and preserve the attribute of the target, the reflection of the face region is also swapped which causes an unnatural result. This abnormal lighting information is usually not captured by existing detection methods, which might be effective assistant information for deepfake detection.

To this end, we propose a light-inconsistency based method for deepfake detection, which leverages the lighting information of the face images to assist the detection. In this paper, we first apply the color constancy technique to face image obtain a pre-processed face image and extract the global lighting information caused by illumination and reflection. The lighting information is directly calculated as the differences between the pre-processed images and the original ones. The lighting information is then fed into the backbone network directly as an assistant channel for training.



(a)Source (b)Target (c)Swapped

Fig. 1. An illustration of the lighting changes before/after swapping. It can be obviously observed that the swapping method can not handle the lighting refinement problem and leaves obvious lighting inconsistency between the swapped face (preserve the lighting information of source) and the target face.

To demonstrate the effectiveness of our lighting inconsistency based method, we conduct extensive experiments on different existing datasets, including FaceForensics++ [16], Celeb-DF [17] and DFDC [24]. It shows that our method is superior to the baseline method. The visualization results show that our method also has good interpretability.

The major contributions in this paper are summarized as follows:

We propose a lighting inconsistency based deepfake detection model, which utilizes the semantic lighting information as assistance to promote the detection ability.

We analyze the visualization results of the lighting differences between the original faces and forged ones. Our method exhibits good interpretability.

·Extensive experiments demonstrate that our method outperforms baseline methods in various datasets.

2 Related Works

Deepfake is indeed a hot research problem in computer vision and graphics recently. Existing deepfake forgery methods try to produce more realistic human faces. This trend makes the deepfake detection methods focus on capturing the subtle visual artifacts of the manipulated faces rather than leveraging more global information, e.g. lighting information. In this section, we will briefly introduce the development of deepfake forgery and detection methods. Besides, we introduce the color constancy which is related to the global lighting information.

2.1 Deepfake Forgery

Rapid progress in deep generative models (such as GAN [1] and VAE [2]) has ignited the interests of both academia and industry. Based on the generative techniques, deepfake forgery achieves tremendous success in very recent years. Currently deepfake forgery methods can be roughly divided into two types: face swapping and face attribute editing. The face swapping methods usually swap the face identity between the source and the target, while the face attribute editing focuses on modifying part of the facial attributes such as transferring the facial expressions and poses from on portrait to another. Face2face [25] is a typical face attribute editing method, which captures the face expression from a source image and transfers it to the target image. DeepFakes [26] and FaceSwap-GAN [27] are two deep generative models based face swapping methods. Recently, FaceShifter [28] proposes a two-stage framework for high fidelity and occlusion-aware face swapping. And Head2Head [29] uses a 3D modeling for specific portrait video to make the poses and expressions controllable for facial reenactment.

Based on those well-developed deepfake forgery methods, many datasets have been proposed to promote the deepfake detection methods, such as Face Forensics++ [16], Celeb-DF [17] and DFDC [24]. However, as we mentioned above, these methods are proposed for more high-fidelity deepfake videos, which usually pay attention to fix the subtle visual artifacts while ignore the global lighting harmonization. Recently, AOT [33] firstly notice the importance of lighting in generating more-realistic results. It adopts the appearance optimal transport model to fix the appearance gaps of illuminations and skin colors between the source and target portraits during the identity swapping. But it has not been widely used in the current deepfake datasets.

2.2 Deepfake Forgery

Since the deepfake forgery has potential societal security concerns, it is of paramount importance to develop effective countermeasures against it. Many works [9–12, 30–32, 34] have been proposed. Early works exploit visual biological artifacts, which is handcrafted, such as eye blinking [10], inconsistent head poses [9], some facial expression changes. XceptionNet [16] is directly used for extracting the spatial features for deepfake detection. Due to the effectiveness, XceptionNet has been a most adopted network architecture in deepfake detection methods. Recently, different perspectives has been considered in emerging works. Multi-attention [19] firstly introduce the idea of fine-grained classification task into deepfake detection and obtained excellent performances

on specific datasets. Face X-ray observed the blending boundary artifacts generated by deepfake forgery and achieves the state-of-art performance on transferability at that time. Spatial Phase Shallow Learning is also a brand new work which focusing on the frequency domain of the manipulated pictures, which achieves a convincing performance on transferability. However, almost all the methods are simply focusing on the texture artifacts, while the semantic information of light is often neglected. As we mentioned in Sect. 2.1, the existing deepfake generation methods perform bad at the semantic level of lighting, so we believe that the semantic information of lighting can be useful auxiliary information for deepfake detection task.

2.3 Color Constancy

Color constancy is also known as white balance, reflects the human brain's judgment of color. Human beings have a psychological tendency not to change the color judgment of a specific object due to light source or external environmental factors, which is called color constancy. Due to the characteristic of color constancy, the forged human face image can also show good realistic to human eyes even the lighting may be very different from the original image. However, the changes in lighting could be a useful information for deepfake detection.

Color constancy has been investigated for decades and numerous conventional algorithms are based on low-level imagery statistics, such as White-Patch, Gray-World, Gray-Edge, Shades-of-Gray, Bright Pixels, Grey Pixel, Gray Index and some other enhancement algorithms.

Recently, there are also some network-based methods in color-constancy, however, conventional methods have been able to accomplish this task well. So in this paper we use Automatic Color Equalization Algorithm (ACE) as the color constancy method.

3 Light-Inconsistency Learning

In this section we will introduce the main structure of our Light-Inconsistency method. We will talk about the light extraction module and main structure in Sect. 3.1, and we will make a further analysis of light inconsistency in Sect. 3.2.

3.1 Main Structure

In common deepfake generating methods, there is no limiting conditions to reconstruct the illumination texture on the manipulated faces, so introducing the lights into the classification models is essential to make it perform better. But how to obtain a method which is fast and effective seems to be a problem. In the field of color constancy in computer vision, the key is to separate the change in the color due to background lighting from the single object itself. In this way, colour constancy is a way to clear out the special filter caused by illumination and reflection, which is what we desire in this task.

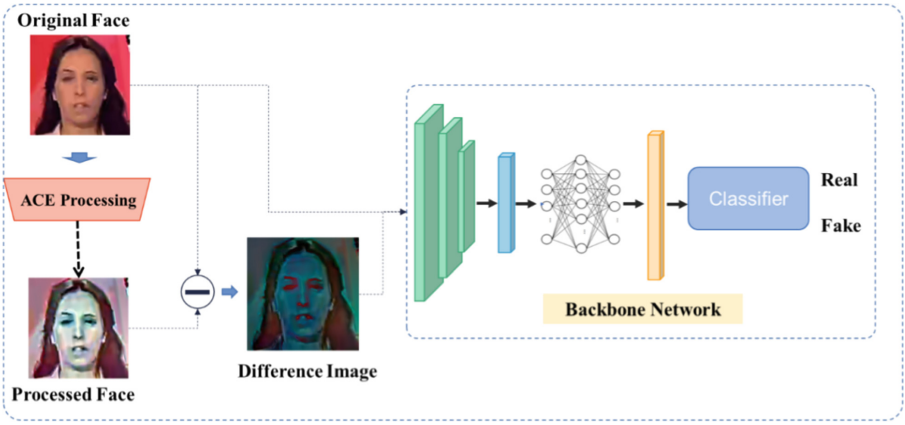


Fig. 2. Main Structure of Light-Inconsistency-Learning. Firstly, we use color constancy methods to obtain a processed image, then an absolute difference is calculated. Both difference and the original input image will be send into the classification model.

Colour constancy methods embraces a considerable variety, including learning-based methods and classical statistics-based methods. Learning-based methods perform better on the single colour constancy task, however, this advantage is basically based on the angular errors, which is not important in deepfake detection. On the other hand, classical statistics-based methods calculate much less than learning-based methods and is much easier to transplant to traditional deepfake detection frameworks. To this end, we use ACE algorithms as our light extraction method.

Ace algorithm is derived from Retinex algorithm. The algorithm considers the spatial position relationship of color and brightness in the image, carries out adaptive filtering of local characteristics, realizes image brightness and color adjustment and contrast adjustment with local and nonlinear characteristics.

Firstly, we adjust the color in spatial domain of the image, complete the color difference correction of the image, and obtain the spatial domain reconstructed image (Fig. 2):

$$R_c(p) = \sum_{j \in Subset, j \neq p} \frac{r(I_c(p) - I_c(j))}{d(p,j)} \tag{1}$$

In this part, R_c Is the intermediate result, $I_c(p) - I_c(j)$ is the brightness difference between two different points, $d(p,j)$ represents the distance measurement function, $r(*)$ is the brightness expression function, which should be an odd function; This step can adapt to the local image contrast, $r(*)$ can enlarge small differences, enrich large differences, and expand or compress the dynamic range according to the local content. Generally, $r(*)$ can be denoted as:

$$r(x) = \begin{cases} 1, & x < -T \\ x/T, & -T \leq x \leq T \\ -1, & x > T \end{cases} \tag{2}$$

In the second part, we need to dynamically expand the corrected image. Ace algorithm is for a single color channel. For color RGB images, each channel needs to be processed separately. Among them, a simple linear extension can be denoted as:

$$\begin{aligned} O_c(p) &= round[127.5 + s_c R_c(P)] \\ M_c &= \max[R_c(p)] \\ m_c &= \min[R_c(p)] \end{aligned} \tag{3}$$

Through the above operations, ACE can be regarded as a simplified model of human visual system, and its enhancement process is consistent with human perception. And this operation could be calculated non-learning and quickly, so it is quite efficient to be used in deepfake detection.

Once the processing procedure was done, we get two inputs for the classification model. One is the original image, and the other one is the absolute difference obtained by calculating the difference between the processed image and the original input. To this end, we obtained a new input image which can be considered as the light information of the input image.

Then we make a channel-wise concatenation before all the images are sent into the classification model. To make things go on, the first layer of the particular classification model must be adjusted to 6 instead of the default number of 3, which means a 3-channel RGB information is added when the classification model is training and testing. And the total 6-channel information is sent into the further layer for convolution and feature with light inconsistency is then extracted. We have to mention that this structure is so simple that it may fit all kinds of neural networks, so this structure is easy to be popularized to all kinds of training-based deepfake detection methods. We can also consider this part as a kind of data augmentation which focus more on the light information which is often neglected by common detecting methods. To this end, we successfully introduced the light information to the training of the deepfake detection network.

Introducing the light information into the classification model is to make the classification model catch the subtle inconsistency and other difference better, which significantly improves the classification accuracy in nearly every task. All the experimental results will be introduced in Sect. 4.

3.2 Analysis of Light Inconsistency

It is obvious that we can catch some of the illumination differences through direct watching the real and synthesized pair of pictures (Fig. 1), but how can we catch more of the subtle lighting difference still remains a question. In this part we will make a further analysis of this certain problem.

We introduced a color-based visual enhancement algorithm to make the visual difference much more evident to observe. The algorithm is listed below:

Algorithm 1 enhancement algorithm

input: *input image*

output: *enhanced image*

function ENHANCEMENT(*RGB input*)
  $result \leftarrow 0$
  for *every pixel* **do**
    $result.r \leftarrow 255 - input.r * magnification\ factor$
    $result.g \leftarrow 255 - input.g * magnification\ factor$
    $result.b \leftarrow 255 - input.b * magnification\ factor$
  end for
  return *result*
end function

In this part, we make an inversion to the absolute difference calculated by the ACE algorithm, and then we make a color enhancement to it. As it is shown in Fig. 3, it is easy to observe that in the third column which is the enhanced difference, there is a obvious difference between the original faces and the manipulated faces. Especially the red part on the cheek of the faces, there is much less red parts in the manipulated faces. The more color in the third column here means that the less part is eliminated during the light extraction processing, and that means, the manipulated faces are more smooth in illumination.

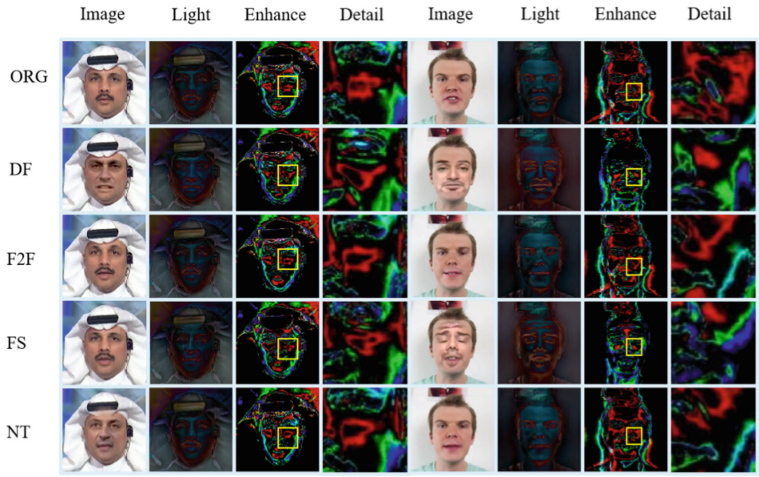


Fig. 3. The analysis of the light inconsistency learning. The first column is the original input picture, the second column is the absolute difference, the third column is the enhancement result, and the fourth is the detail of the enhanced difference. The obvious difference is boxed.

4 Experiments

In this section, we present extensive experiments on different datasets to demonstrate the effectiveness of our approach.We test the performances of our method and two baseline methods on FaceForensics++ [16], Celeb-DF [17] and DFDC [24], which are three commonly used datasets for deepfake detection.

4.1 Evaluations on FF++

In this section, we first compare our method to a baseline method XceptionNet on FF++ [16]. We evaluate our methods on different video compression settings including high quality (HQ (c23)) and low quality(LQ(c40)). Table 1 shows the testing accuracy comparison of the original XceptionNet our method based on XceptionNet. Our method obtains an obvious improvement or is on par with baseline on both HQ and LQ settings.

It is worth mentioning that the certain improvement is huge on the specific task on NeuralTextures, which is a task that embraces less texture-based artifacts. This represents our method has a stronger ability to capture semantic artifacts than simply using network-based methods. This can also be seen on LQ dataset results that the improvement is way higher than the improvement on HQ datasets. Low-quality videos have been compressed strongly so common neural networks can not capture the texture-based features well. The improvement of accuracy performance mainly benefits from the extra light-inconsistency learning capturing the semantic artifacts.

Since our method is backbone independent, it is easy to transfer the lighting inconsistency learning strategy to other backbones. In Table 2, we apply our method to ResNet-50 and compare the performance to the original ResNet-50. Our method can also enhances a convincing improvement.

Table 1. Comparisons to XceptionNet on FF++

Datasets	DF_c23	NT_c23	F2F_c23	FS_c23	Total_c23
XceptionNet	98.69	93.67	**98.7**	**99.08**	92.83
Ours	**98.92**	**94.37**	98.34	97.75	**95.74**
Datasets	DF_c40	NT_c40	F2F_c40	FS_c40	Total_c40
XceptionNet	95.21	78.02	90.66	91.99	84.75
Ours	**97.41**	**80.37**	**90.68**	**92.93**	**85.52**

Table 2. Comparisons to ResNet-50 on FF++

Datasets	DF_c23	NT_c23	F2F_c23	FS_c23	Total_c23
ResNet-50	97.55	91.33	**98.25**	**98.25**	92.06
Ours	**97.68**	**92.16**	98.14	97.53	**95.41**
Datasets	DF_c40	NT_c40	F2F_c40	FS_c40	Total_c40
ResNet-50	96.42	76.58	88.36	88.36	84.11
Ours	**97.26**	**78.65**	**88.92**	**90.84**	**85.47**

4.2 Evaluations of Celeb-DF and DFDC

To further demonstrate the effectiveness of our methods. We also compare our method to the baseline method on Celeb-DF and DFDC datasets. Both datasets have better

visual-quality videos compare to FF++. DFDC is the most challenging deepfake dataset since the fake videos in it are quite hard to distinguish by human eyes. Obvi-ously, the results in Table 3 demonstrate that our method also achieves better per-formances on such challenging datasets.

Table 3. Comparisons to XceptionNet on DFDC and Celab-DF

Datasets	DFDC	Celab-DF
XceptionNet	78.26	98.24
Ours	**80.47**	**98.96**

5 Conclusion

In this work, we propose a brand new face forgery detection method focusing on the light information which is often neglected by common detection methods. The core competence of our work is that the light information contains more abundant semantic feature and these feature will help our classification model perform better. Besides, our work forces the network to focus more on the light inconsistency of the manipulated faces to capture semantic artifacts for more robustness. We use extensive experiments to prove that our method can make an obvious improvement on the face forgery detection, especially on the low quality tasks.

References

1. Goodfellow, I.J., et al.: Generative adversarial networks. arXiv:1406.2661 (2014)
2. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv:1312.6114 (2013)
3. Huh, M., Liu, A., Owens, A., Efros, A.A.: Fighting fake news: image splice detection via learned self-consistency. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision–ECCV 2018. ECCV 2018. LNCS, vol. 11215, pp. 101–117. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01252-6_7
4. Jeong, Y., et al. DoFNet: depth of field difference learning for detecting image forgery. In: Ishikawa, H., Liu, C.L., Pajdla, T., Shi, J. (eds.) Computer Vision–ACCV 2020. ACCV 2020. LNCS, vol 12627. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-69544-6_6
5. Kwon, P., You, J., Nam, G., Park, S., Chae, G.: Kodf: a large-scale korean deepfake detection dataset. arXiv:2103.10094 (2021)
6. Lee, S., Tariq, S., Shin, Y., Woo, S.S.: Detecting handcrafted facial image manipulations and GAN-generated facial images using shallow-FakeFaceNet. Appl. Soft Comput. **105**, 107256 (2021)
7. Nguyen, T.T., Cuong, M., Nguyen, D.T., Nguyen, D.T., Nguyen, S., Saeid, N.: Deep learning for deepfakes creation and detection. arXiv:1909.11573 (2019)
8. Sun, K., Liu, H., Ye, Q., Liu, J., Gao, Y., Shao, L.: Domain general face forgery detection by learning to weight (2021)

9. Yang, Y.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)
10. 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 (2018)
11. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Two-stream neural networks for tampered face detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1831–1839 (2017)
12. Xu, B., Liu, J., Liang, J., Lu, W., Zhang, Y.: Deepfake videos detection based on texture features. Comput. Mater. Contin. **68**(1), 1375–1388 (2021)
13. 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 (2018)
14. Huy, H., Nguyen, J., Yamagishi, J., Echizen, I.: Capsule-forensics: using capsule networks to detect forged images and videos. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2307–2311 (2019)
15. Li, S.: Exposing deepfake videos by detecting face warping artifacts. arXiv:1811.00656 (2018)
16. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Niener, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1–11 (2019)
17. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: a largescale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3207–3216 (2020)
18. 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)
19. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., Yu, N.: Multiattentional deepfake detection. arXiv:2103.02406 (2021)
20. Liu, Z., Qi, X., Torr, P.H.: Global texture enhancement for fake face detection in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8060–8069 (2020)
21. Sun, X., Wu, B., Chen, W.: Identifying invariant texture violation for robust deepfake detection. arXiv:2012.10580 (2020)
22. Chugh, K., Gupta, P., Dhall, A., Subramanian, R.: Not made for each other-audio-visual dissonance-based deepfake detection and localization. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 439–447 (2020)
23. Zhang, X., Karaman, S., Chang, S.F.: Detecting and simulating artifacts in gan fake images. In: 2019 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6 (2019)
24. Dolhansky, B., Howes, R., Pflflaum, B., Baram, N., Ferrer, C.C.: The deepfake detection challenge (dfdc) preview dataset. arXiv:1910.08854 (2019)
25. Thies, J., Zollhöfer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2Face: real-time face capture and reenactment of RGB videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2387–2395 (2016)
26. DeepFakes (2017). http://github.com/ondyari/FaceForensics/tree/master/dataset/
27. FaceSwap (2017). http://github.com/deepfakes/faceswap
28. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Faceshifter: towards high fidelity and occlusion aware face swapping. arXiv:1912.13457 (2019)
29. Koujan, M.R., Doukas, M.C., Roussos, A., Zafeiriou, S.: Head2head: video-based neural head synthesis. arXiv:2005.10954 (2020)

30. Baumy, A., Algarni, A.D., Abdalla, M., El-Shafai, W.E.F.: Efficient forgery detection approaches for digital color images. Comput. Mater. Contin. **71**(2), 3257–3276 (2022)
31. Tan, W., Wu, Y., Wu, P., Chen, B.: A survey on digital image copy-move forgery localization using passive techniques. J. New Media **1**(1), 11–25 (2019)
32. Munawar, M., Noreen, I.: Duplicate frame video forgery detection using Siamese-based RNN. Intell. Autom. Soft Comput. **29**(3), 927–937 (2021)
33. Zhu, H., Fu, C., Wu, Q.: AOT: Appearance optimal transport based identity swapping for forgery detection. In: NeurIPS (2020)
34. Zhou, P., Han, X., Morariu, V., Davis, L.: Two stream neural networks for tampered face detection. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1831–1839 (2017)