Deepfakes Catcher: A Novel Fused Truncated DenseNet Model for Deepfakes Detection

Fatima Khalid, Ali Javed, Aun Irtaza, and Khalid Mahmood Malik

Abstract In recent years, we have witnessed a tremendous evolution in generative adversarial networks resulting in the creation of much realistic fake multimedia content termed deepfakes. The deepfakes are created by superimposing one person's real facial features, expressions, or lip movements onto another one. Apart from the benefits of deepfakes, it has been largely misused to propagate disinformation about influential persons like celebrities, politicians, etc. Since the deepfakes are created using different generative algorithms and involve much realism, thus it is a challenging task to detect them. Existing deepfakes detection methods have shown lower performance on forged videos that are generated using different algorithms, as well as videos that are of low resolution, compressed, or computationally more complex. To counter these issues, we propose a novel fused truncated DenseNet121 model for deepfakes videos detection. We employ transfer learning to reduce the resources and improve effectiveness, truncation to reduce the parameters and model size, and feature fusion to strengthen the representation by capturing more distinct traits of the input video. Our fused truncated DenseNet model lowers the DenseNet121 parameters count from 8.5 to 0.5 million. This makes our model more effective and lightweight that can be deployed in portable devices for real-time deepfakes detection. Our proposed model can reliably detect various types of deepfakes as well as deepfakes of different generative methods. We evaluated our model on two diverse datasets: a large-scale FaceForensics (FF)++ dataset and the World Leaders (WL) dataset. Our model achieves a remarkable accuracy of 99.03% on the WL dataset and 87.76% on the FF++ which shows the effectiveness of our method for deepfakes detection.

F. Khalid · A. Irtaza
Department of Computer Science, University of Engineering and Technology, Taxila, Pakistan

A. Javed (✉)
Department of Software Engineering, University of Engineering and Technology, Taxila, Pakistan
e-mail: ali.javed@uettaxila.edu.pk

K. M. Malik
Department of Computer Science and Engineering, Oakland University, Rochester, MI, USA

S. Anwar et al. (eds.), *Proceedings of International Conference on Information Technology and Applications*, Lecture Notes in Networks and Systems 614,
https://doi.org/10.1007/978-981-19-9331-2_20

Keywords Deepfakes detection · DenseNet121 · FaceForensics++ · Fused truncated DenseNet · World leaders dataset

1 Introduction

The evolution of deep learning-based algorithms such as autoencoders [12] and Generative Adversarial Networks (GANs) [9] have led to the generation of many realistic image and video-based deepfakes. Deepfakes represent the synthesized multimedia content based on artificial intelligence which mainly falls in the categories of FaceSwap, Lip-Sync, and Puppet mastery. FaceSwap deepfakes are centered on identity manipulation, where the original identity is swapped with the targeted one. Lip-syncing is a technique for modifying a video where the mouth area fits the arbitrary audio, whereas the puppet-mastery approach is concerned with the modification of facial expressions including the head and eye movement of the person. Deepfakes videos have some useful applications such as creating videos of a deceased person by using his single photo, changing the aging and de-aging of people, etc. Both applications can also be used to create realistic videos of live and deceased actors in the entertainment industry. Deepfakes have the potential not only to influence our view of reality, but can also be used for retaliation and deception purposes by targeting politicians and famous leaders and spreading disinformation to take political revenge.

Existing literature on face-swapping and puppet-mastery has explored different end-to-end deep learning (DL)-based approaches. Various studies [3, 5, 10, 11] have focused on the application of DL-based methods for face swap deepfakes detection. In Bonettini et al. [3] ensemble model of EfficientNet and average voting was proposed. The model was evaluated only in intra-dataset settings, thus the generalization capability of this method cannot be guaranteed in an inter-dataset setup. In Rossler et al. [11], CNN was used in conjunction with the SVM for real and face swap detection. This approach was unable to perform well on the compressed videos. In Nirkin et al. [10] confidence score was computed from cropped faces, which were later fed into the deep learner to identify the identity manipulation. This model does not generalize well on unseen data. In de Lima et al. [5], VGG-11 was used to determine frame level features, which were then fed to various models like ResNet, R3D, and I3D to detect the real and forged videos. This technique is computationally more costly.

Research approaches [1, 4, 6, 14] have also been presented for puppet mastery deepfakes detection by employing the DL-based methods. In Guo et al. [6], feature maps generated from convolutional layers were subtracted from the original images. The method removes unnecessary details from the image, allowing the RNN to concentrate on the important details. This method requires more samples for training to obtain satisfactory performance. In Zhao et al. [14], pairwise learning was used to extract source features from CNN, which were later used for classification. However, the performance of the model decreases on images that have consistent features. In Chintha et al. [4], temporal discrepancies in deepfakes videos were identified by combining XceptionNet CNN which extracted the facial features using bidirectional LSTM. The architecture performed well on multiple datasets; however, the performance degrades on compressed samples. In Agarwal et al. [1], a combination of VGG-16 and encoder-decoder network was applied for detection by computing the facial and behavioral attributes. This method does not apply well to unseen deepfake videos.

According to the literature, existing approaches, notably [1, 10], don't have the generalization ability on the unseen data. Rossler et al. [11], Chintha et al. [4] performs well on high-quality videos, but their performance degrades on compressed videos. Although [5] outperforms other state-of-the-art (SOTA) techniques, but is computationally complex. To better address the challenges, we present a novel fused truncated DenseNet framework that works effectively on unseen data and induces modifications to further reduce the computational cost and optimization efforts while achieving higher accuracy. Specifically, this paper makes a significant contribution based on the following:

1. We present a novel fused truncated DenseNet model that is robust to different types of deepfakes (face swap, puppet mastery, imposter, and lip-sync) and to different generative methods (Face2Face, NeuralTextures, Deepfakes, and FaceShifter).
2. We present an efficient deepfakes detection method by employing the GeLu activation function in our proposed method to reduce the complexity of the model.
3. We introduce a series of layers including global average pooling and dense layers combined with the regularization technique to prevent overfitting.
4. To evaluate the generalizability of our proposed model, we performed extensive experiments on two different deepfakes datasets including the cross-set examination.

2 Proposed Methodology

This section explains the proposed workflow employed for deepfakes detection. The architecture of our proposed framework is depicted in Fig. 1.

2.1 Facial Frames Extraction

The initial stage is to identify and extract faces from the video frames since the facial landmarks are the most manipulated part in deepfakes videos. For this purpose, we used the Multi-task Cascaded Convolutional Networks (MTCNN) [15] face detector to extract the facial region of 300 × 300 from the input video during pre-processing. This approach recognizes the facial landmarks such as the eyes, nose, and mouth,

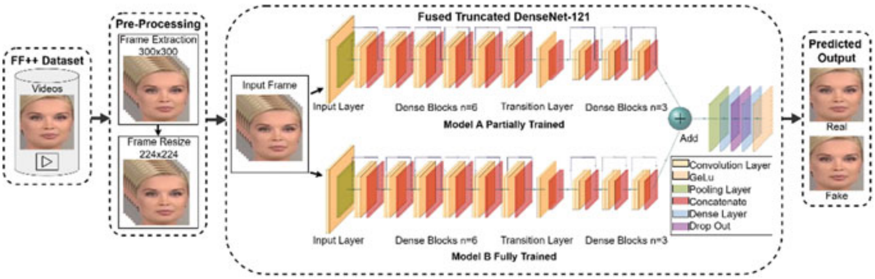

Fig. 1 Architecture of proposed method

from coarse to fine details. We chose this method as it detects faces accurately even in the presence of occlusion and variable light, unlike other face detectors such as Haar Cascade and Viola Jones framework [13].

2.2 Fused Truncated DenseNet

We extract the frames having frontal face exposure after detecting faces in the input video. The frames were then resized to 224 × 224 resolution and fed to our fused truncated DenseNet121. We introduce truncation modifications that help in parameter and model size minimization; as well as feature fusion, which merges the correlated feature values produced by our algorithm. As a result, an effective and lightweight model for the detection of real and deepfake videos is created. The use of pre-trained frameworks is inspired by the fact that these models have been trained on enormous publicly available datasets like ImageNet, and hence can learn the essential feature points.

DenseNet121 is a ResNet architectural extension. The training technique faces vanishing gradient issues as the network's depth grows. Both the ResNet and DenseNet models are intended to address this issue. The DenseNet design is built on all layer's connectivity, with each layer receiving input from all previous layers and passing the output to all the subsequent layers. As a result, the resultant connections are dense, which enhances the efficiency with fewer parameters. The goal of having a DenseNet121 model is to give a perfect transmission of features throughout the whole network without performance degradation, even with considerable depth. DenseNet also handles parameter inflation utilizing a concatenation instead of layer additions.

Our proposed method includes two DenseNet121 architectures. Model A is partially trained on our dataset, its early layers are frozen to preserve the ImageNet features, and the remaining layers are retrained on our data. Whereas model B is entirely retrained on our dataset. Figure 1 shows the proposed fused truncated DenseNet model, which is composed of 7 × 7 Convolution layer, proceeded by the Batch Normalization (BN), Gaussian Error Linear Unit (GeLu), and 3 × 3 Max

Pooling layer. Next, a pair of dense blocks with a BN, GeLu, and 1 × 1 Convolution layer is followed by another BN, GeLu, and 1 × 1 Convolution layer. Unlike ResNet and other deep networks that rely on feature summation to generate large parameters, the DenseNet model employs a dense block with '*n*' rate of growth that is appended to all the network layers. This approach evolves into an efficient end-to-end transfer of features from preceding layers to succeeding layers. The proposed design produces a rich gradient quality even at deeper depths while lowering the parameter count makes it very useful for detection purposes. To avoid depletion of resources during the features extraction, the DenseNet model needs a transition layer that down-samples the feature maps by using 1 × 1 Convolution layer and 2 × 2 Average Pooling layer.

Layer Truncation Although DenseNet has much lesser parameters than other DL-based models, the proposed approach aims to further minimize the parameters without compromising its effectiveness. DenseNet121 has around 8.5 million parameters. The base DenseNet model is suitable for large datasets such as the ImageNet, which has over 14 million images and 1000 categories, training and replicating this model can be time-consuming. Furthermore, with such a small dataset, employing the complete model's architecture merely adds complexity and uses enormous resources. As a result, most of the models' layers are eliminated through a proposed truncation from its complete network, lowering the number of parameters and reducing the end-to-end flow of features. The proposed fused truncated DenseNet with only six dense blocks followed by a transition layer connecting to another set of three dense blocks is shown in Fig. 1. The proposed methodology reduces the DenseNet121 model's parameter count by a significant factor of 93.5%. More specifically, truncated DenseNet decreases the parameters from the initial 8.5 million to only around half a million.

Activation Function is used in a multilayer neural network to express the connection between the output values of neurons in the preceding layer and the input values of those in the following layer. It determines whether a neuron should be activated or not. We used the Gaussian Error Linear Units (GeLu) [7] function in our method. As sigmoid and ReLu faces the gradient vanishing issue, along with this, ReLu also creates the dead ReLu issue. To address these issues of ReLu, probabilistic regularization techniques such as dropout are widely used after the activation functions to improve accuracy. GeLu is presented to combine stochastic regularization with an activation function. It is a conventional Gaussian distribution function that puts nonlinearity to the output of a neuron depending on their values, rather than using the input value as in ReLu.

Model concatenation and Prediction The smaller size of the truncated DenseNet network results in a lower parameter value. On the contrary, adding more depth to the layers will make the truncation approach useless. To overcome this problem, we employed the model concatenation method, which improves the accuracy of our model with fewer parameters. Model concatenation and feature fusion broadened the model instead of increasing its depth, enabling the required fast end-to-end feature extraction for training and validation. To better process the features produced by the fusion of both models, the proposed method incorporates a new set of layers consisting of Global Average Pooling, a dense layer, and the dropout connected to another dense layer activated by the classifier. These additional layers attempt to increase efficiency and regularization, hence preventing overfitting problems.

3 Experiment Setup and Results

3.1 Dataset

We evaluated the performance of the proposed method using two datasets: FaceForensics++ [11] and the World Leaders Dataset [2]. FF++ is an extensive face manipulation dataset created with automated, modern video editing techniques. Two traditional computer graphics methods, Face2Face (F2F) and FaceSwap (FS) are used in conjunction with two learning-based methods, DeepFakes (DF) and NeuralTextures (NT). Each video has an individual with a non-occlusion face, although it is difficult due to differences in the skin tone of various people, lighting conditions, the presence of facial accessories, and the loss of information due to low video resolution.

The YouTube videos of world-famous politicians (Clinton, Obama, Warren, and others) with their original, comical imposters (Imp), face swap (FS), lip-sync (LS), and puppet master subsets made up the WL dataset. Politicians are speaking throughout the videos; each video has only one person's face and the camera is static with minimal variations in zooming. We divided both datasets into 80:20 splits with 80% of the videos for training and the rest 20% for testing.

3.2 Performance Evaluation of the Proposed Method

We designed an experiment to analyze the performance of our method on the original and fake sets of FF++ and WL datasets to demonstrate its effectiveness for deepfakes detection. For this purpose, we employed our model to classify the real and fake videos of each subset of FF++ separately. On FF++, we tested the real samples with the fake samples from FS, DF, F2F, NT, and FaceShifter (FSh) sets, and the results are presented in Table 1. It can be noticed that the FF++-FS set has the highest accuracy of 95.73% and 0.99 AUC among all other sets. FS videos are generated by using the 3D blending method. These remarkable results on the FS set indicate that our model can better capture these traits to identify the identity changes and static textures. Whereas FSh achieved an accuracy of only 60.90% and AUC of 0.67 because the generative method of this set is very complex due to the fusion of two complex GAN's architecture [8]. This makes it extremely challenging to reliably capture the distinctive traits of the texture used in the FSh, which limits the accuracy of our model.

Table 1 Performance evaluation of proposed method on FF++ dataset

	FS	DF	F2F	NT	FSh
Accuracy	95.73	93.9	92.6	83.5	60.90
PR	0.99	0.97	0.97	0.90	0.63
AUC	0.99	0.98	0.97	0.92	0.67

Table 2 Performance evaluation of proposed method on WL dataset

Leaders	Subsets	Accuracy	PR	AUC
Obama	FS	94.57	0.96	0.97
	Imp	58.57	0.60	0.63
	LS	62.36	0.65	0.68
JB	FS	89.68	0.91	0.94
	Imp	95.65	0.97	0.96
Clinton	FS	84.13	0.87	0.86
	Imp	91.43	0.92	0.94
Warren	FS	93.14	0.93	0.95
	Imp	93.12	0.93	0.95
Sander	FS	89.59	0.91	0.90
	Imp	78.88	0.80	0.82
Trump	Imp	99.70	1.00	1.00

For WL, each leader's deepfakes type (FS, Imp, and LS) is tested with the original samples. Table 2 shows that the FS of Obama has shown the best accuracy of 94.57% and 0.97 AUC. Whereas Imp set of Trump has shown an accuracy of 99.70% among all the leaders. The results of this experiment revealed that our proposed model performed remarkably on both datasets. These results are due to the GeLu's nonlinear behavior and its combinative property of dropout, zoneout, and ReLu. GeLu solves the dying ReLu problem by providing a gradient in the negative axes to prevent neurons from dying and is also capable of differentiating each datapoint of the input image.

3.3 Ablation Study

In this experiment, an ablation study is conducted to demonstrate the performance of various activation functions on the FaceSwap set of the FaceForensics++ dataset. Table 3 illustrates the performance of different activation functions. The results show that our method employing the GeLu activation provided the best performance as compared to other activation functions. The disparity in findings is mainly due to the GeLu's combinative property of dropout and zone out as well as its non-convex,

Table 3 Performance evaluation on different activation functions

Activation functions	ReLu	SeLu	TRelu	ELU	GeLu
Testing on FF++ (FS)	94.5	90.6	92.3	95.09	95.73

non-monotonic, and nonlinear nature with curvature present in all directions. On the other hand, convex and monotonic activations like ReLu, ELU, and SeLu are linear in the positive axes and lack curvature. As a result, GeLu outperforms other activation functions.

3.4 Performance Evaluation of Proposed Method on Cross-Set

In this experiment, we designed a cross-set evaluation to inspect the generalizability of the proposed method among the intra-sets of the datasets. For the FF++ dataset, we conducted an experiment where each trained set is tested on all the other sets, like FS trained set is tested on all the other sets, respectively. Similarly, for the WL dataset, we conducted the same experiment within each leader's intra-set, like Obama's FS trained set is tested on the Imp and LS sets, respectively. The results displayed in Table 4 are slightly encouraging as both the datasets contain different deepfakes types and generative methods, but still our proposed method can differentiate the modifications of identity change, expression change, and neural rendering.

Table 4 shows that, on the FF++ dataset, the sets having the same generative method achieved better results as compared to others. In comparison to the FF++ dataset, our proposed model has shown better results on the WL dataset, it has easily detected the FS and Imp of most of the leaders with good accuracies, as both the types have the same generative methods, so our model generalizes well on the same generative methods. LS of Obama has shown the lowest accuracies among all because this set contains spatiotemporal glitches. DL-based models (CNNs along with RNNs) can extract the features in both the spatial and temporal domains. In our method, we used a fused truncated DenseNet-based CNN model to identify the artifacts in the spatial domain only, which reduces the accuracy of this set.

We conducted another cross-set evaluation experiment for the WL dataset, where the FS and Imp trained model of one leader is tested with the FS and Imp of another leader, respectively. The motive behind this experiment was to check the robustness of the same forgery type on different leaders. The results shown in Table 5 are relatively good, which shows that the proposed model can distinguish the same forgery on different individuals even in the presence of challenging conditions such as variations in skin tones, facial occlusions, lightning conditions, and facial artifacts.

Table 4 Performance evaluation on cross-sets of FF++ and WL dataset

Test set										
Train set	FF++	Subsets		FS	DF	F2F	FSh	NT	Imp	LS
		FS		–	48.6	67.0	52.9	49.2	–	–
		DF		51.9	–	54.8	58.1	68.9	–	–
		F2F		51.4	54.7	–	50.2	57.0	–	–
		FSh		56.0	56.0	51.8	–	48.1	–	–
		NT		55.2	55.2	50.2	48.3	–	–	–
	WL	Obama	FS	–	–	–	–	–	62.1	46.9
			Imp	48.0	–	–	–	–	–	32.2
			LS	35.3	–	–	–	–	41.2	–
		JB	FS	–	–	–	–	–	76.0	0.94
			Imp	79.2	–	–	–	–	–	–
		Clinton	FS	–	–	–	–	–	84.8	–
			Imp	83.2	–	–	–	–	–	–
		Warren	FS	–	–	–	–	–	82.1	–
			Imp	92.0	–	–	–	–	–	–
		Sander	FS	–	–	–	–	–	76	–
			Imp	91.0	–	–	–	–	–	–

Table 5 Performance evaluation on cross-set of WL dataset

Test set												
		Obama		JB		Clinton		Warren		Sander		Trump
		Fs	Imp	Fs	Imp	Fs	Imp	Fs	Imp	Fs	Imp	Imp
Train set	Obama	–	–	66.3	60.3	71.3	55.1	53.8	50.1	50.3	79.3	71.1
	JB	69.3	53.6	–	–	37.6	84.1	69.4	71.2	87.3	75.2	69.3
	Clinton	65.1	59.1	81.0	70.2	–	–	60.4	62.8	81.1	61.2	55.2
	Warren	78.4	48.3	83.1	61.1	71.3	80.1	–	–	91.2	70.1	69.4
	Sander	65.1	42.2	79.6	65.0	84.3	61.2	51.4	49.1	–	–	79.2
	Trump	–	51.3	–	75.2	–	68.2	–	55.5	–	82.1	–

3.5 Comparative Analysis with Contemporary Methods

The key purpose of this experiment is to validate the efficacy of the proposed model over existing methods. The performance of our method on the FF++ with existing methods is shown in Table 6. The accuracy of our model for FS and NT has increased by 5.44 and 2.9%, respectively. Whereas, for F2F and DF, our method has achieved higher accuracies as compared to most of the methods. It is difficult to obtain good

Table 6 Performance comparison against existing methods on FF++ dataset

Model	FS	DF	F2F	NT	FSh	Combined
XceptionNet	70.87	74.5	75.9	73.3	–	62.40
Steg. Features	68.93	73.6	73.7	63.3	–	51.80
ResidualNet	73.79	85.4	67.8	78.0	–	55.20
CNN	56.31	85.4	64.2	60.0	–	58.10
MesoNet	61.17	87.2	56.2	40.6	–	66.00
XceptionNet	90.29	96.3	86.8	80.6	–	70.10
Classification	54.07	52.3	92.77	–	–	83.71
Segmentation	34.04	70.37	90.27	–	–	93.01
Meso-4	–	96.9	95.3	–	–	–
MesoInception	–	98.4	95.3	–	–	–
Proposed	95.73	93.9	92.6	83.5	60.9	87.76

Table 7 Performance comparison against existing methods on WL dataset

Paper	Subset	Obama	Clinton	Warren	Sander	Trump	JB	Combined
Agarwal et al. [2]	FS	0.95	0.95	0.98	0.96	–	–	0.93
	Imp	0.94	0.93	1.00	0.94	0.94	–	
	LS	0.83	–	–	–	–	–	
Agarwal et al. [1]	–	–	–	–	–	–	–	0.94
Proposed	FS	0.97	0.86	0.95	0.90	–	0.94	0.97
	Imp	0.63	0.94	0.95	0.82	1.00	0.96	
	LS	0.68	–	–	–	–	–	

detection results on all subsets of the FF++ dataset, especially in the presence of challenging conditions like non-facial frames, varying illumination conditions, people of different races, and the presence of facial accessories. Our method outperforms most methods since it achieves good identification results across all subsets and can discriminate between real and fake videos generated using different manipulation techniques.

We compared the performance of our method on the WL dataset with existing methods using the AUC score. Table 7 shows when all the dataset's leaders are combined, our method outperforms the existing techniques.

4 Conclusion

In this paper, we have presented a fused truncated DenseNet model to better distinguish between real and deepfakes videos. Our proposed system is lightweight and resilient with a shorter end-to-end architecture and fewer parameter sizes. In comparison to other SOTA models with greater parameter sizes, our truncated model trains quicker and performs well on a large and diverse dataset. Our model performed well, regardless of the distinct occlusion settings, variations in skin tones of people, and the presence of facial artifacts in both datasets. We performed an intra-set evaluation on both datasets and get better results on the sets having the same type of generative method. This shows that our model can detect the deepfakes on the unseen samples of any dataset using similar generative methods for deepfake creation. We intend to increase the generalizability of our methodology in the future to improve the cross-corpus assessment.

Acknowledgements This work was supported by the grant of the Punjab Higher Education Commission of Pakistan with Award No. (PHEC/ARA/PIRCA/20527/21).

References

1. Agarwal S, Farid H, El-Gaaly T, Lim S-N (2020) Detecting deep-fake videos from appearance and behavior. In: 2020 IEEE international workshop on information forensics and security (WIFS)
2. Agarwal S, Farid H, Gu Y, He M, Nagano K, Li H (2019) Protecting world leaders against deep fakes. CVPR workshops
3. Bonettini N, Cannas ED, Mandelli S, Bondi L, Bestagini P, Tubaro S (2021) Video face manipulation detection through ensemble of cnns. In: 2020 25th international conference on pattern recognition (ICPR)
4. Chintha A, Thai B, Sohrawardi SJ, Bhatt K, Hickerson A, Wright M, Ptucha R (2020) Recurrent convolutional structures for audio spoof and video deepfake detection. IEEE J Sel Top Sign Proces 14(5):1024–1037
5. de Lima O, Franklin S, Basu S, Karwoski B, George A (2020) Deepfake detection using spatiotemporal convolutional networks. arXiv:2006.14749
6. Guo Z, Yang G, Chen J, Sun X (2021) Fake face detection via adaptive manipulation traces extraction network. Comput Vis Image Underst 204:103170
7. Hendrycks D, Gimpel K (2016) Gaussian error linear units (gelus). arXiv:1606.08415
8. Li L, Bao J, Yang H, Chen D, Wen F (2019) Faceshifter: towards high fidelity and occlusion aware face swapping. arXiv:1912.13457
9. Liu M-Y, Huang X, Yu J, Wang T-C, Mallya A (2021) Generative adversarial networks for image and video synthesis: algorithms and applications. Proc IEEE 109(5):839–862
10. Nirkin Y, Wolf L, Keller Y, Hassner T (2021) DeepFake detection based on discrepancies between faces and their context. IEEE Trans Pattern Anal Mach Intell
11. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2019) Faceforensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF international conference on computer vision
12. Tewari A, Zollhoefer M, Bernard F, Garrido P, Kim H, Perez P, Theobalt C (2018) High-fidelity monocular face reconstruction based on an unsupervised model-based face autoencoder. IEEE Trans Pattern Anal Mach Intell 42(2):357–370
13. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001

14. Zhao T, Xu X, Xu M, Ding H, Xiong Y, Xia W (2021) Learning self-consistency for deepfake detection. In: Proceedings of the IEEE/cvf international conference on computer vision, pp 15023–15033

15. Xiang J, Zhu G (2017) Joint face detection and facial expression recognition with MTCNN. In: 2017 4th international conference on information science and control engineering (ICISCE). IEEE, pp 424–427