symmetry

Article

Deepfake Video Detection Based on MesoNet with Preprocessing Module

Zhiming Xia [1], Tong Qiao [1,2,*], Ming Xu [1,3,*], Xiaoshuai Wu [1], Li Han [4] and Yunzhi Chen [3]

[1] School of Cyberspace, Hangzhou Dianzi University, Hangzhou 310005, China; pcmg@hdu.edu.cn (Z.X.); shinewu@hdu.edu.cn (X.W.)
[2] State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou Science and Technology Institute, Zhengzhou 450064, China
[3] School of Information Engineering, Hangzhou Vocational & Technical College, Hangzhou 310018, China; 2006010022@hzvtc.edu.cn
[4] School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310005, China; hanli@hdu.edu.cn
* Correspondence: tong.qiao@hdu.edu.cn (T.Q.); mxu@hdu.edu.cn (M.X.)

Abstract: With the development of computer hardware and deep learning, face manipulation videos represented by Deepfake have been widely spread on social media. From the perspective of symmetry, many forensics methods have been raised, while most detection performance might drop under compression attacks. To solve this robustness issue, this paper proposes a Deepfake video detection method based on MesoNet with preprocessing module. First, the preprocessing module is established to preprocess the cropped face images, which increases the discrimination among multi-color channels. Next, the preprocessed images are fed into the classic MesoNet. The detection performance of proposed method is verified on two datasets; the AUC on FaceForensics++ can reach 0.974, and it can reach 0.943 on Celeb-DF which is better than the current methods. More importantly, even in the case of heavy compression, the detection rate can still be more than 88%.

Keywords: multimedia forensics; Deepfake detection; preprocessing module; MesoNet

check for updates

Citation: Xia, Z.; Qiao, T.; Xu, M.; Wu, X.; Han, L.; Chen, Y. Deepfake Video Detection Based on MesoNet with Preprocessing Module. *Symmetry* **2022**, *14*, 939. https://doi.org/10.3390/sym14050939

Academic Editors: José Carlos R. Alcantud and Zichi Wang

Received: 29 March 2022
Accepted: 26 April 2022
Published: 5 May 2022

1. Introduction

Traditional face tampering is a tedious and time-consuming process, which requires professional video editing tools and professional knowledge. With the continuous development of computer hardware and deep learning, image synthesis has been an enormous breakthrough. A vast amount of face manipulation videos such as Deepfake have spread on video sharing sites and social media (see Figure 1), "seeing is not always believing" is being achieved to some extent. As the advance of generative adversarial networks (GAN [1]) and Auto-Encoders, these techniques are increasingly being applied to Deepfake, enabling the creation and rapid distribution of high-quality video tampering content.

While Deepfake technology has advanced society in some ways, there are concerns about the harmful effects of its misuse [2]. As Deepfake requires only a small number of face photos to enable video face-swapping, some malicious users have taken advantage of the data available on the internet to generate numerous fake videos. The porn industry is the first to embrace this technology, where many face-swapping porn videos are spread on the internet featuring female celebrities. Replacing the heroine of pornographic movies with female stars and forging some video content for politicians, company executives, and other influential people to achieve misleading public opinion, winning selection, and manipulating stock prices. Many deceptive face-swapping videos pose a substantial potential threat to national security, social stability, and personal privacy [3]. Verifying the authenticity of videos disseminated online is gradually becoming one of the hot topics in digital society.

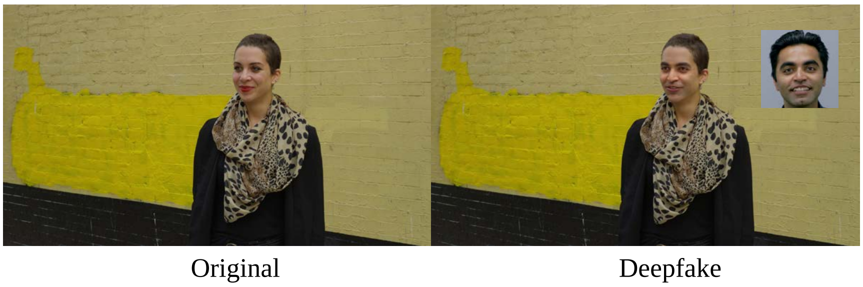


Figure 1. Example of Deepfake on social media, the image from DeepFakeDetection (DFD) dataset.

Nowadays, most videos or images on social media are compressed. For example, when users use social software such as Twitter or Facebook to send videos, the size of the uploaded video is limited, and the user must compress it before uploading. If criminals distribute compressed Deepfake videos, it would be difficult for us to detect them effectively. To address the problem, the forensics of compressed Deepfake videos becomes a meaningful and challenging task. Many Deepfake detection methods have been proposed and achieved good detection performance. However, from the perspective of symmetry, the formidable challenges for Deepfake detection still exist: when under compression attacks, the detection performance of most detectors dramatically decreases due to the loss of image feature information. In addition, most methods do not perform well for the new Deepfake algorithms. Therefore, improving the detector's performance in complex scenes with highly deceptive tampered videos, is urgently need. In light of the aforementioned challenges, this paper proposes a discriminator based on MesoNet with preprocessing module. The main contributions of this paper are as follows:

- We propose a new preprocessing module to filter low-frequency signals in images and retain high-frequency signals, and therefore increasing the discrimination between Deepfake generated and real images. The effectiveness of the preprocessing module is verified in the ablation experiment.
- We propose a new Deepfake detection method by combining the classic MesoNet with preprocessing module. In the case of heavy compression, it can still maintain good robustness; the accuracy of our proposed method is still higher than 88%.
- The performance of our method is verified among numerous baseline datasets. Extensive experimental evaluations demonstrate that the proposed method performs well on Celeb-DF and FaceForensics++. Our method outperforms some SOTA methods on the Celeb-DF. In addition, the AUC of our proposed method is 0.965 and 0.843 on the UADFV and DFD datasets, respectively.

The rest of this paper is organized as follows: Section 2 comprehensively summarizes the state of the arts. Section 3 introduces the detection method proposed in this paper. Section 4 illustrates the numerical results on the benchmark datasets. Section 5 concludes this paper.

2. Related Works

In this section, we provides an introduction to Deepfake detection methods. The current detection methods can be divided into two categories according to different feature extraction methods: Deepfake video detection methods based on hand-crafted and Deepfake video detection methods based on deep learning.

2.1. Deepfake Video Detection Method Based on Hand-Crafted

The Deepfake video detection methods based on hand-crafted features mainly relies on specific tampering traces. Detection is achieved by extracting frequency domain features and statistical features with significant differences in images, such as the photo response non-uniformity (PRNU [4]), image quality assessment, optical flow [5], etc. In synthesis and post-processing, some unique image feature information is often left in the video. Therefore, some traditional image forensics methods can be used for Deepfake video detection. For such detection methods, researchers need to manually extract feature information with more excellent discrimination from face images, and feed the obtained feature information into the classifier for training, and finally get a detection model.

Due to the significant difference in the generation mechanism between Deepfake and real videos, Koopman et al. [6] took the lead in proposing to use PRNU to detect Deepfake videos. They cropped the face region of each frame, divided it into eight groups, and calculated the average PRNU value of each group. The experimental results showed a significant difference in the average normalized correlation coefficient between the real video and the Deepfake video. Lugstein et al. [7] also proposed to use PRNU to detect Deepfake videos and achieved good detection results. Frank et al. [8] believe that PRNU is effective in detecting Deepfake videos on small datasets. When the detection model trained based on PRNU features is tested on the image dataset generated by GAN, the detection results are not very convincing. Faten et al. [9] used edge features for detection, extracted image texture feature information through various image feature point detectors, and then sent the obtained features into SVM for training. On the small-scale dataset UADFV , the accuracy of the detection model is over 90%.

Although the forensics technology based on traditional image features is very mature and can achieve good detection results, when dealing with some, it generates poor quality Deepfake videos. When the forged content is processed in the face of compression, blurring, and noise processing, the original image features would be destroyed. The detection model trained based on such features is easily bypassed, and the detection rate will drop.

2.2. Deepfake Video Detection Method Based on Deep Learning

The Deepfake video detection method based on deep learning relies on the powerful learning ability of neural networks and the increasingly rich sample set. These methods are further divided into two categories: frame-level detection methods and video-level detection methods. The frame-level detection methods mainly focus on the image feature information in a single frame, and design different network structures for feature analysis. The video-level detection methods focus on intra-frame information and consider inter-frame temporal features.

2.2.1. Frame-Level Detection Methods

Zhou et al. [10] propose a two-stream network, using a GoogleNet to analyze face tampering artifacts and a patch-based triplet network to analyze the steganalysis features, respectively. Similarly, Afchar et al. [11] propose Meso-4 and MesoInception-4 networks. Deepfake detection is performed with the help of image mesoscopic features. The method is trained and tested on the Deepfake dataset constructed by the author and has achieved good detection results. Rossler et al. [12] used the Xception network to detect Deepfake videos on the FF++ dataset. In addition, Li et al. [13] introduced the Spatial Pyramid Pooling layer based on ResNet-50. This module can better focus on the change of the resolution of the face area. Hu et al. [14] propose a frame inference-based detection framework (FInfer) to solve the problem of high visual quality Deepfake video detection. By predicting the facial representation of future frames from the facial representation of the current frame, a prediction loss is designed to maximize the discriminability of real and Deepfake videos. Nirkin et al. [15] propose to use Two-Networks to detect Deepfake videos by detecting the different signal features of face-swapping regions. The method involves two networks: a face recognition network that considers face regions bounded by tight semantic segmentation and a context recognition network that considers face context (e.g., hair, ears, neck). Most detection methods model Deepfake detection as a general binary classification problem. Since the difference between real and fake images in this task is usually subtle and local, Zhao el al. [16] treat Deepfake detection as a fine-grained classification problem and propose a novel multi-attention Deepfake detection network.

2.2.2. Video-Level Detection Methods

Although convolutional neural networks have powerful image analysis and feature extraction capabilities, these methods often only focus on intra-frame feature information and ignore some inter-frame features. Compared with still images, the video contains rich inter-frame details. Considering the temporal continuity between frames, Guera et al. [17] first propose using Recurrent Neural Networks (RNNs) to detect Deepfake videos. Since the face images in the fake videos are generated frame by frame, the Auto-Encoders does not consider the previous images when generating, which often leads to multiple anomalies in some areas between the previous and subsequent frames. Inspired by [17], Sabir et al. [18] propose to combine temporal network and face preprocessing techniques to detect tampered videos. The video will destroy the continuity between adjacent frames, resulting in the above detection method not performing well under compression. Monsterrat et al. [19] propose a CNN-RNN network based on an automatic weighting mechanism to solve the above problem. The framework increases the weight of face images during video-level detection to improve the robustness of the detection algorithm. The experimental results show that it is feasible and efficient to learn image-specific feature information through the network.

Since the detection performance of the methods would be significantly decreased in the face of compression attacks, Hu et al. [20] propose a dual-stream network detection method to increase the robustness. The dual-stream networks analyze the intra-frame noise features and the inter-frame temporal features, respectively, and prune the network to prevent overfitting. Good detection performance has been achieved on FF++ and Celeb-DF datasets. Agarwal et al. [21] found that the facial action units of different people are inconsistent, and the "micro-expression" feature is used to detect Deepfake videos. In their experiment, 19 facial expression units and head action units are selected. The presence intensity of the target individual's facial expression unit and head action unit in the video are extracted through Openface2. The obtained presence intensity is then used as a feature set to feed the SVM classifier for training. Although this detection method is robust to compression attacks and noise addition attacks, it is limited to specific individuals. Wu et al. [22] propose a new Deepfake detection method to solve the above problem. The method detects tampered faces through combining spatial features and steganalytic features, and obtains good generalization ability. Masi et al. [23] also propose a two-branch network to improve the generalization of the detection method. One branch network is used to extract original image feature information, and the other branch is used to suppress facial content. It can achieve good performance when cross-tested on various datasets.

Detection methods based on deep learning are prone to overfitting in the model training stage. Hence, researchers exploit the inherent differences between real and fake videos through preprocessing.

3. Proposed Method

In this section, we first use the preprocessing module to preprocess the face image, then input the processed image to the baseline network for training, and finally the detection model outputs the predicted probability of each frame. The pipeline of our proposed Deepfake detection method is shown in Figure 2.

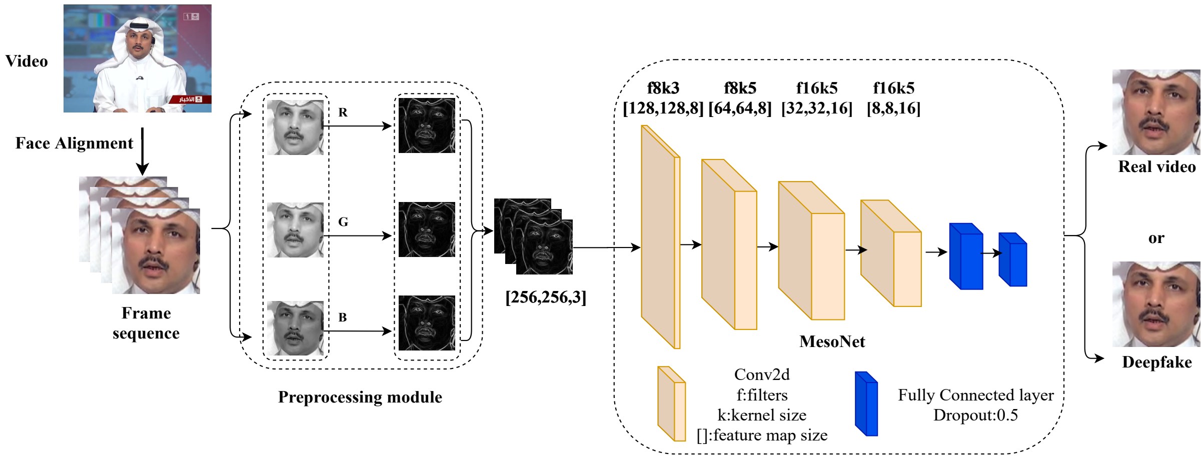


Figure 2. The pipeline of our proposed Deepfake detection method.

3.1. Preprocessing Module

Compared with the real images, the images generated by Deepfake are fundamentally different in the imaging mechanism. When Auto-Encoders is generating an image, the encoder learns the pixel statistical distribution of the input image. Then the decoder reconstructs the face image in the corresponding mode [24]. We observe that the face regions of target individuals are smoother in Deepfake videos compared to real videos. An image can be roughly divided into textured and smooth regions, where textured regions represent potentially high variations in pixel values. The high texture area of the image contains rich feature information, and the feature information can guide us through the detection tasks.

Taking this as a clue, we devote our efforts to reduce the influence of smooth regions on detection. Therefore, we consider using a preprocessing module to augment such texture differences. The preprocessing module can be divided into two parts: face cropping to obtain face image and face image preprocessing, since only the face area in the Deepfake video has been tampered with. To reduce the background area's influence on subsequent detection, we use Dlib to crop out the face region. It is observed that the face area in the Deepfake video is generally smooth. Through a new preprocessing method, the low-frequency information of the face image is filtered, and the high-frequency information with high texture discrimination is retained, thereby improving the detection performance of the detection model. Inspired by Qiao et al. [25], we designed a new preprocessing method to process the face image. The specific definition of preprocessing method is as follows:

$$p_{i,j} = \sqrt{(m_{i,j} - m_{i+1,j})^2 + (m_{i,j} - m_{i,j+1})^2} \qquad (1)$$

where $i \in \{1, \ldots, H\}$, $j \in \{1, \ldots, W\}$, $m_{i,j}$ are the pixel values, $p_{i,j}$ is the pixel value after preprocessing method, H and W represent the size of the image.

The pixel values of the preprocessed face image are calculated from the adjacent pixel values of the original image. In the smooth area of the image, the difference between these pixel values is small, so the pixel value of the smooth area tends to be zero after preprocessing method. In high-texture areas, the difference between these pixel values is large, and the high texture area would still be retained after preprocessing. Therefore, after preprocessing module, the texture information can be largely preserved. As shown in Figure 3, the first row displays the normal and corresponding forged face images, and the images in the second row are the image processed by the preprocessing method.

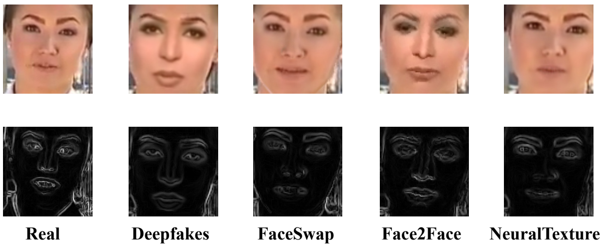


Figure 3. The real, Deepfakes, FaceSwap, Face2Face, NeuralTexture image, and corresponding preprocessed image on the FF++.

This paper mainly investigates the difference on three color channels of R, G, and B. It is proposed to use the chi-square distance to measure the differences between real images and Deepfake generated images. For a given set of real images, we first calculate the variance value of each image, and construct its corresponding statistical histogram. Similarly, for a given set of Deepfake generated images, we construct another statistical histogram in the same way. Finally, the chi-square distance of the constructed histogram is calculated to estimate the similarity between the real face image set and the Deepfake generated face image set.

To verify whether the difference between real and fake images can be increased after preprocessing module, we calculate the chi-square distance after preprocessing module on FF++ and Celeb-DF datasets. The experimental results are shown in Figure 4 . It can be observed that after the preprocessing module, the differences in facial texture are significantly increased, especially on the FF++ dataset. The chi-square distance on the R, G, and B channels are 1.039, 1.169, and 1.208. The differences have also increased on the Celeb-DF dataset, referring to 0.578, 0.513, and 0.712, respectively. We use t-SNE [26] to visualize the preprocessed statistical variance value of each image. The Figure 5 shows the statistical variance value without preprocessing module. It can be seen from the Figure 5 that there is less discrepancy between the real and fake images without the preprocessing module. In contrast, as shown in the Figure 6, the difference between the images is increased after the preprocessing module. Therefore, through the above statistical analysis, it is assumed that preprocessing module can increase the difference. Notably, we also perform ablation experiments in Section 4 to verify the effectiveness of the preprocessing module.

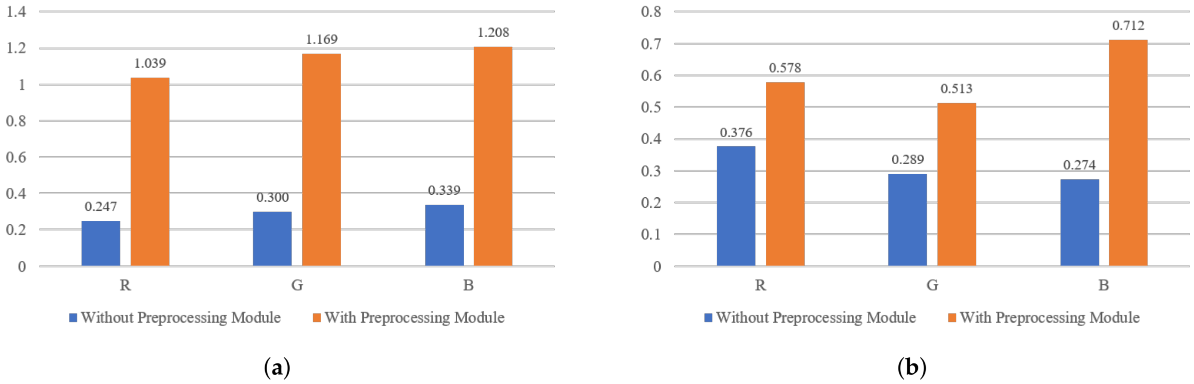
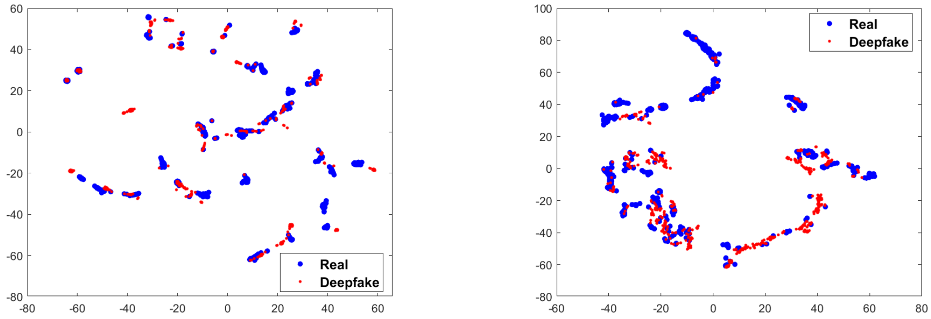


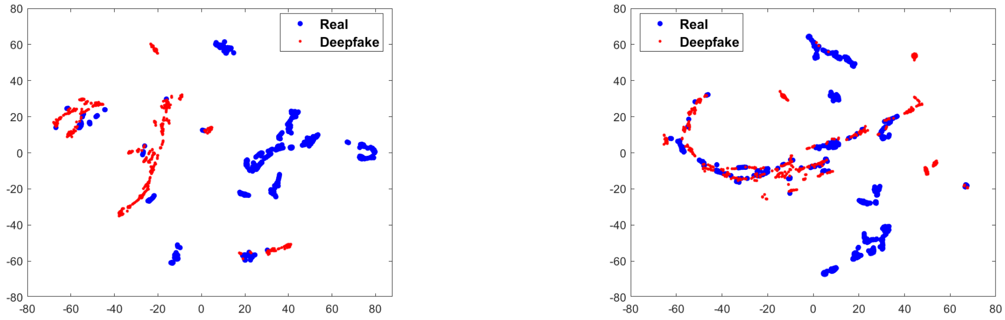
Figure 4. Comparison of chi-square distance on FF++ and Celeb-DF. (**a**) Comparison of chi-square distance on FF++ dataset; (**b**) Comparison of chi-square distance on Celeb-DF dataset.



(**a**)

(**b**)

Figure 5. Illustration of t-SNE visualization of statistical variance value without the preprocessing module on FF++ and Celeb-DF. (**a**) 2-D visualization on FF++ dataset; (**b**) 2-D visualization on Celeb-DF dataset.



(**a**)

(**b**)

Figure 6. Illustration of t-SNE visualization of statistical variance value after the preprocessing module on FF++ and Celeb-DF. (**a**) 2-D visualization on FF++ dataset; (**b**) 2-D visualization on Celeb-DF dataset.

3.2. MesoNet

When the video or image content is compressed, the performance of Deepfake detection methods based on low-level noise feature will significantly decrease. This is because some feature information will be destroyed when the image is compressed, and the noise information that the detector relies on will also be attenuated. Similarly, at the image semantic level, especially when the image is describing a face, the network cannot effectively distinguish between authentic face images and Deepfake generated images. Therefore, Archfar et al. [11] propose MesoNet based on the Inception module. MesoNet is a neural network with few neural layers that mainly focuses on image mesoscopic features, which are between high-level and low-level features.

Archfar et al. [11] first conducted experiments on a more complex CNN network structure and then gradually simplified the network structure. It should be noted that the detection capability of the simplified network structure is comparable to that of the complex network structure. Finally, the proposed MesoNet, is mainly composed of 4 consecutive convolutional layers and two fully connected layers. The network starts with four consecutive convolutions and pooling, followed by a dense layer as a hidden layer. A nonlinear ReLu activation function is used in each convolutional layer to improve generalization. Batch Normalization is used for regularization to avoid vanishing gradient effects. And the Dropout layer is used to regularize each fully connected layer to enhance the network's robustness. The network has a total of 27,977 trainable parameters. The specific network details are shown in Figure 7.

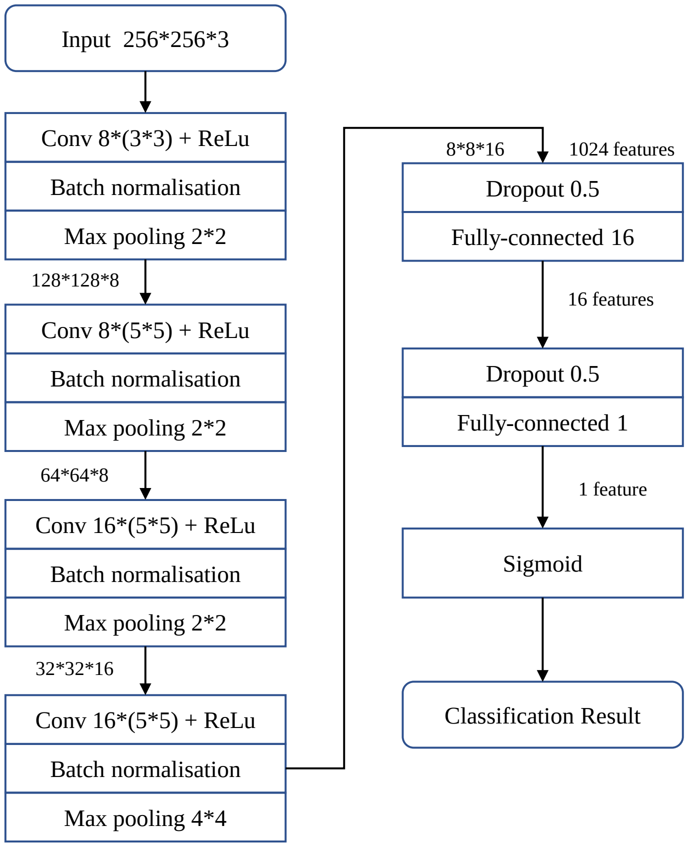


Figure 7. The network architecture of MesoNet [11].

4. Experiments

In this section, we first introduce two benchmark datasets, and meanwhile elaborating experimental settings. Next, we perform ablation experiments to verify the effectiveness of the preprocessing module. Then, we validate our proposed method on the FaceForensics++ [12] and Celeb-DF [13] datasets. The comparison experiments with SOTA are also demonstrated. In addition, it is proposed to confront different types of Deepfake videos. Finally, the robustness performance facing compression attacks is verified.

4.1. Dataset

In this context, we use two public datasets, FF++ and Celeb-DF. The FF++ dataset contains different scenarios to simulate real forensics. These include 1000 original videos containing only a single face, concerning news broadcasts, exclusive interviews, one-person talk shows, downloaded from Youtube. Based on the original video, each video is tampered with the four techniques of Deepfakes, Face2Face, FaceSwap, and NeuralTexture. Additionally, three compressed versions (c0, c23 and c40) are available on the FF++ to simulate realistic scenarios such as social network platforms.

The Celeb-DF dataset is proposed by Li et al. [13], which belongs to the second-generation Deepfake video dataset. This dataset mainly aims at solving the problems of low-quality generated faces, obvious splicing boundaries, and rough tampering traces in the first-generation datasets such as UADFV [27], FaceForensics++, Deepfake-TIMIT [28]. The Celeb-DF dataset improves the Deepfake generation algorithm, increases the size of the generated image in the model training phase, and increases the brightness and contrast of the face image to reduce the inconsistency between the tampered area and the surrounding area. As a result, the tampered videos in the dataset are more deceptive, and have higher visual quality. Moreover, the Celeb-DF dataset, released in 2020, consists of 590 real videos and 5639 high-quality Deepfake videos with an average length of 13 s and the frame rate is 30.

4.2. Experiment Settings

All videos in the datasets use Dlib 19.14.0 for face landmarks calibration and cropping. Since the diverse sizes of cropped face images, it is necessary to uniformly resize the face images to 256 × 256 to complete the following experiments. The experiments are run on an Ubuntu 18.04 system, using 4 × 11 GB Nvidia GeForce GTX 1080ti for model training. All deep learning models are implemented on the Keras 2.2.4 framework. Due to a large number of similar frames in the video, to reduce the data scale, 20 frames are randomly selected from each video in the datasets. The image set is further divided into training sets, validation set and test set in a ratio of 8:1:1. The images are randomly chosen. The relevant parameters of model training are as follows: batch size is 64; the training epoch is 200; the initial learning rate is 0.00001; the decay rate is 0.00005. The cross-entropy loss is chosen as the loss function, and SGD solver is used for optimization. Moreover, the shuffle variable during training is set to True. The performance evaluation of the experiment is AUC and ACC.

4.3. Ablation

We conjecture the texture difference can be increased using preprocessing module (see Figure 4). It is proposed to perform ablation experiments on the FF++ and Celeb-DF datasets to verify the effectiveness of the preprocessing module. Table 1 illustrates the results on the FF++ dataset and the Celeb-DF dataset. The detectors with preprocessing module outperform the detectors without preprocessing module. Meanwhile, our proposed method achieves good detection results on both datasets. The highest AUC on the FF++ dataset is 0.974, and the highest AUC on the Celeb-DF dataset reaches 0.943. Therefore, the preprocessing module can significantly improve the performance of the detector.

Table 1. The accuracy and AUC of our proposed method for each color channel on FaceForensics++ and Celeb-DF.

Method	Color Channel	FaceForensics++		Celeb-DF	
		ACC	AUC	ACC	AUC
MesoNet	R	0.912	0.922	0.711	0.831
	G	0.884	0.891	0.645	0.747
	B	0.893	0.917	0.755	0.834
	RGB	0.885	0.902	0.731	0.837
MesoNet + Preprocessing module	R	**0.941**	**0.974**	0.936	0.931
	G	0.896	0.932	0.941	0.942
	B	0.939	0.969	**0.949**	**0.943**
	RGB	0.916	0.912	0.899	0.884

4.4. Comparison with Previous Methods

To further verify the performance of our proposed method, this subsection compares the proposed method with current arts. Among them, Two-Networks [15], Xception [12], and DefakeHop [29] are the latest image-based detection methods. Capsule [30] , FWA [31], DSP-FWA [13] are the earlier deep learning based detection methods, and HeadPose [27], VA-MLP, and VA-LogReg [32] are methods based on hand-crafted features. Also, some novel approaches have aroused attention such as DeepfakeUCL [33], Robust-GAN [34] and Triplets [35]. Moreover, we adopt the baseline MesoNet network [11] with preprocessing module for detection comparison in the R, G, and B color channels.

The frame level comparison results are listed in Figure 8. It is worth noting that the Xception is tested with their pre-trained model, and the FWA and DSP-FWA are only trained on the FF++ dataset. On the FF++ dataset, the performance of the proposed method rivals that of the state-of-the-art detectors based on deep learning. In addition, the performance of the proposed method is superior to detection methods based on hand-crafted features. Our proposed detection method based on MesoNet with preprocessing module achieves the best ACC and AUC of 0.941 and 0.974 on the FF++ dataset. Figures 9 and 10 show ROC in the frame level. Most of the current detection methods can achieve good results in the FF++ dataset. The main reason is that many defects appear in the FF++ dataset. With these artifacts, good detection performances can be obtained. For example, Capsule [30] mainly focuses on the position of facial features in the image. Due to the poor quality generated in the FF++ dataset, the facial features of the target characters in some videos are not coordinated. Hence, this method can achieve excellent performance. In the new generation dataset represented by Celeb-DF, the performance of our proposed method is better than the most current state-of-the-art methods. The highest AUC of our proposed method can reach 0.943, and the lowest AUC is 0.884. In addition, Robust-GAN [34] incorporates the attention mechanism into the improved Xception network and fuses the feature information of RGB and YCbCr color channels to achieve detection. It can reach 0.759 on the Celeb-DF dataset. Detection methods based on hand-crafted features like HeadPose, VA-MLP, and VA-LogReg rely on specific tampering traces, such as uncoordinated head poses and uncoordinated head poses. Such defects are corrected in the new generation of deepfake videos, so these three methods are less effective in the new dataset.

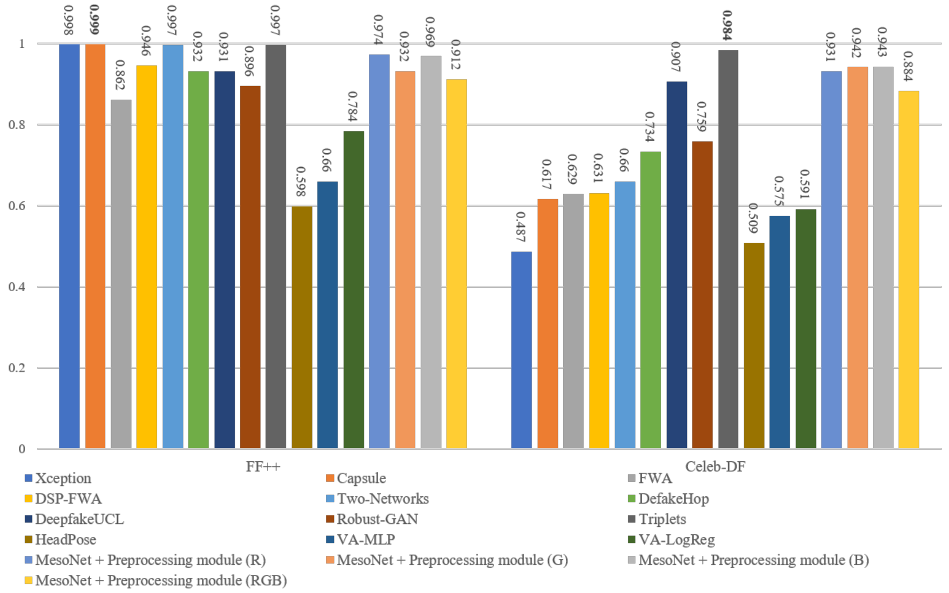


Figure 8. The frame level AUC results with SOTA on the FF++ and Celeb-DF datasets.

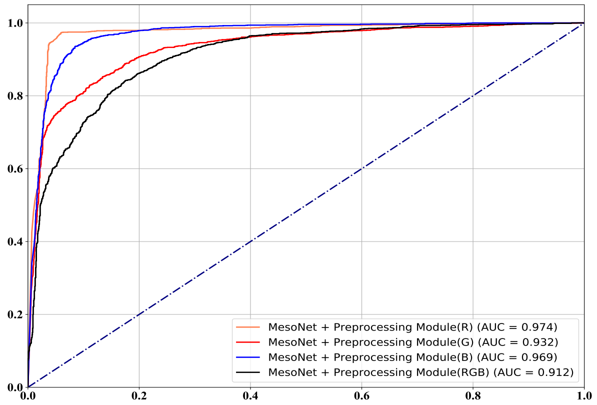


Figure 9. The ROC on FaceForensics++.

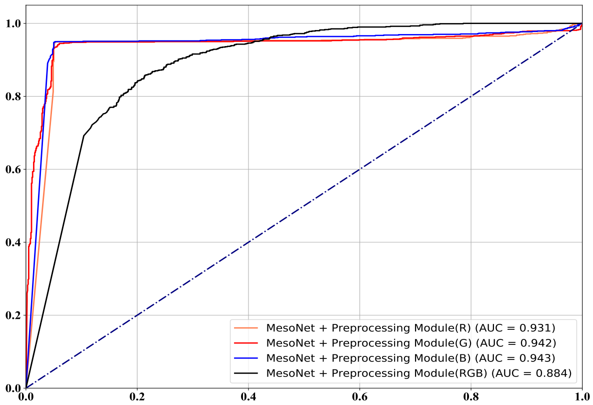


Figure 10. The ROC on Celeb-DF.

Furthermore, considering the differences in imaging mechanisms between real and Deepfake generated images, we conduct experiments on three single-color channels of R, G, and B, respectively. The experimental results are also listed in Figure 8. The experimental results show that our method performs better on three single channels than on RGB three-channels.

4.5. Detecting Different Types of Deepfake Videos

There are two main types of tampering in Deepfake videos: facial identity tampering such as *Deepfakes* and facial expression tampering such as *Face2Face*. Therefore, it is necessary to evaluate whether the currently proposed method can deal with different types of Deepfake videos with multiple types of tampering. We conduct the experiments on the FF++ dataset, and the experimental results are shown in Table 2. Our proposed method can achieve good results in these four tampering methods. As a result, the AUC reached 0.974, 0.959, 0.974 and 0.963, respectively. Therefore, the proposed method can effectively resist against different Deepfake videos.

Table 2. The performance of our method on different types of Deepfake videos.

Type	**R**		**G**		**B**	
	ACC	**AUC**	**ACC**	**AUC**	**ACC**	**AUC**
Deepfakes	0.941	**0.974**	0.896	0.932	0.939	0.969
FaceSwap	0.951	**0.959**	0.849	0.914	0.857	0.923
Face2Face	0.969	**0.974**	0.885	0.956	0.865	0.939
NeuralTexture	0.921	0.942	0.937	0.960	0.932	**0.963**

In fact, different datasets usually use different Deepfake generation algorithms. To measure the performance of our method on other datasets, we conduct performance validation on UADFV [36] and DeepFakeDetection(DFD) [37] datasets. The UADFV dataset belongs to the first-generation dataset, containing 49 real videos and 49 corresponding Deepfake videos. Compared with FF++ and Celeb-DF datasets, DFD contains more scenes (such as kitchen pan, meeting scenarios) to imitate scenarios in the wild. There are even multiple people in some scenes, so the detection task for this dataset is more challenging. We thus evaluate our proposed method performance on UADFV, FF++, Celeb-DF, and DFD datasets. As can be observed in Table 3, the detection results of our proposed method on each dataset are 0.969 for UADFV, 0.974 for FF++, 0.942 for Celeb-DF, and 0.843 for DFD. The detection results on the first-generation datasets (UADFV, FF++) are significantly better than those on the second-generation datasets (Celeb-DF, DFD). The main reason is that the first-generation datasets have relatively obvious artefacts such as obvious splicing boundaries, and the color of the face area and the background area are inconsistent. The face area is smoother, especially in the first-generation dataset, and better results can be obtained. The performance of our proposed method has an unavoidable decrease in the DFD. In contrast, as DFD contains a variety of scenes, and some tampered videos contain multiple faces, which are more deceptive.

Table 3. The AUC of our proposed method for each color channel on UADFV, FF++, Celeb-DF and DFD.

Method	UADFV	FF++	Celeb-DF	DFD
MesoNet+Preprocessing module(R)	0.948	**0.974**	0.931	0.816
MesoNet+Preprocessing module(G)	**0.965**	0.932	0.942	0.828
MesoNet+Preprocessing module(B)	0.952	0.969	**0.943**	**0.843**

4.6. Robustness

Most of the Deepfake videos spreading on social networks are usually compressed. Therefore, it is necessary to verify whether our proposed method can maintain good detection performance under compression attacks. We conduct compression attack experiments on the FF++ dataset. Table 4 shows the experimental results of our method and other methods under different compression factors. As Table 4 illustrates, although the detection rate of our method has a slight declining trend, it still remains above 88%. The main reason is that the preprocessing module can filter out low-frequency signals in the image and increase the difference between two types of videos.

Table 4. The accuracy (%) of our method on three different compressed versions of the FF++ dataset. c0 refers to the video without compression, and c23 refers to slight compression, and c40 refers to heavy compression.

Methods	c0	c23	c40
Steg.Features [38]	97.63%	70.97%	55.98%
MesoNet [11]	95.23%	83.10%	70.47%
Cozzolino et al. [39]	98.57%	78.45%	58.69%
Bayar et al. [40]	**98.75%**	82.97%	66.84%
Rahmouni et al. [41]	97.03%	79.08%	61.18%
MesoNet+Preprocessing module	94.12%	**91.97%**	**88.51%**

5. Conclusions

Nowadays, the threats posed by Deepfake are well known. Many Deepfake video detection methods are therefore proposed, but most of them suffer from remarkable performance degradation in the case of compression attacks. To deal with this problem, in this paper, we propose a Deepfake detection method based on MesoNet with preprocessing module. Specifically, by filtering low-frequency signals in the image, high-frequency signals with apparent differences are retained, thus increasing the texture difference between the authentic and Deepfake generated images. Then the preprocessed image is fed into the MesoNet to learn deep features to complete detection. The effectiveness of the proposed method is verified on benchmark datasets. It is worth noting that our proposed method outperforms other baseline methods on the baseline dataset Celeb-DF. In addition, our proposed method is robust to compression attacks with different factors.

Author Contributions: Conceptualization, Z.X.; Funding acquisition, Z.X.; Investigation, T.Q. and M.X.; Methodology, T.Q.; Software, Y.C.; Writing-original draft, Z.X.; Writing-review and editing, X.W. and H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work was supported by the Fundamental Research Funds for the Provincial Universities of Zhejiang under grant No. GK219909299001-007.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
2. Khichi, M.; Yadav, R.K. A Threat of Deepfakes as a Weapon on Digital Platform and their Detection Methods. In Proceedings of the 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 6–8 July 2021; pp. 1–8.
3. Yu, P.; Xia, Z.; Fei, J.; Lu, Y. A survey on deepfake video detection. *IET Biom.* **2021**, *10*, 607–624. [CrossRef]
4. Lukáš, J.; Fridrich, J.; Goljan, M. Detecting digital image forgeries using sensor pattern noise. In Proceedings of the Security, Steganography, and Watermarking of Multimedia Contents VIII, San Jose, CA, USA, 15 January 2006; Volume 6072, pp. 362–372.
5. Amerini, I.; Galteri, L.; Caldelli, R.; Del Bimbo, A. Deepfake video detection through optical flow based cnn. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019.
6. Koopman, M.; Rodriguez, A.M.; Geradts, Z. Detection of deepfake video manipulation. In Proceedings of the 20th Irish machine vision and image processing conference (IMVIP), Ulster University, Ulster, Northern Ireland, 29–31 August 2018.
7. Lugstein, Florian and Baier, Simon and Bachinger, Gregor and Uhl, Andreas PRNU-based Deepfake Detection. In Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security, Virtual, 22–25 June 2021; pp. 7–12
8. Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; Holz, T. Leveraging frequency analysis for deep fake image recognition. In Proceedings of the International Conference on Machine Learning (PMLR), Virtual, 13–18 July 2020; pp. 3247–3258.
9. Kharbat, F.F.; Elamsy, T.; Mahmoud, A.; Abdullah, R. Image feature detectors for deepfake video detection. In Proceedings of the 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, 3–7 November 2019; pp. 1–4.
10. Zhou, P.; Han, X.; Morariu, V.I.; Davis, L.S. Two-stream neural networks for tampered face detection. In Proceedings of the 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1831–1839.
11. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. Mesonet: A compact facial video forgery detection network. In Proceedings of the 2018 IEEE international workshop on information forensics and security (WIFS), Hong Kong, China, 11–13 December 2018; pp. 1–7.
12. Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 1–11.
13. 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, Seattle, WA, USA, 13–19 June 2020; pp. 3207–3216.
14. Hu, J.; Liao, X.; Liang, J.; Zhou, W.; Qin, Z. FInfer: Frame Inference-based Deepfake Detection for High-Visual-Quality Videos. *IEEE Trans. Pattern Anal. Mach. Intell.* 2022, 1–9. in press.
15. Nirkin, Y.; Wolf, L.; Keller, Y.; Hassner, T. DeepFake detection based on discrepancies between faces and their context. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [CrossRef]
16. Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; Yu, N. Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2185–2194.
17. Güera, D.; Delp, E.J. Deepfake video detection using recurrent neural networks. In Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 27–30 November 2018; pp. 1–6.
18. Sabir, E.; Cheng, J.; Jaiswal, A.; AbdAlmageed, W.; Masi, I.; Natarajan, P. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)* **2019**, *3*, 80–87.
19. Montserrat, D.M.; Hao, H.; Yarlagadda, S.K.; Baireddy, S.; Shao, R.; Horváth, J.; Bartusiak, E.; Yang, J.; Guera, D.; Zhu, F.; et al. Deepfakes detection with automatic face weighting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 668–669.
20. Hu, J.; Liao, X.; Wang, W.; Qin, Z. Detecting Compressed Deepfake Videos in Social Networks Using Frame-Temporality Two-Stream Convolutional Network. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 1089–1102. [CrossRef]

21. Agarwal, S.; Farid, H.; Gu, Y.; He, M.; Nagano, K.; Li, H. Protecting World Leaders Against Deep Fakes. In Proceedings of the CVPR Workshops, Long Beach, CA, USA, 16–21 June 2019; Volume 1.
22. Wu, X.; Xie, Z.; Gao, Y.; Xiao, Y. Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 2952–2956.
23. Masi, I.; Killekar, A.; Mascarenhas, R.M.; Gurudatt, S.P.; AbdAlmageed, W. Two-branch recurrent network for isolating deepfakes in videos. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 667–684.
24. Li, H.; Li, B.; Tan, S.; Huang, J. Identification of Deep Network Generated Images Using Disparities in Color Components. *Signal Process.* **2020**, *174*, 107616. [CrossRef]
25. Qiao, T.; Shi, R.; Luo, X.; Xu, M.; Zheng, N.; Wu, Y. Statistical model-based detector via texture weight map: Application in re-sampling authentication. *IEEE Trans. Multimed.* **2018**, *21*, 1077–1092. [CrossRef]
26. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
27. Yang, X.; Li, Y.; Lyu, S. Exposing deep fakes using inconsistent head poses. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 8261–8265.
28. Korshunov, P.; Marcel, S. Deepfakes: A new threat to face recognition? assessment and detection. *arXiv* **2018**, arXiv:1812.08685.
29. Chen, H.S.; Rouhsedaghat, M.; Ghani, H.; Hu, S.; You, S.; Kuo, C.C.J. DefakeHop: A Light-Weight High-Performance Deepfake Detector. In Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6.
30. Nguyen, H.H.; Yamagishi, J.; Echizen, I. Capsule-forensics: Using capsule networks to detect forged images and videos. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 2307–2311.
31. Li, Y.; Lyu, S. Exposing deepfake videos by detecting face warping artifacts. *arXiv* **2018**, arXiv:1811.00656.
32. Matern, F.; Riess, C.; Stamminger, M. Exploiting visual artifacts to expose deepfakes and face manipulations. In Proceedings of the 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa, HI, USA, 7–11 January 2019; pp. 83–92.
33. Fung, S.; Lu, X.; Zhang, C.; Li, C.T. DeepfakeUCL: Deepfake Detection via Unsupervised Contrastive Learning. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–8.
34. Chen, B.; Liu, X.; Zheng, Y.; Zhao, G.; Shi, Y.Q. A robust GAN-generated face detection method based on dual-color spaces and an improved Xception. *IEEE Trans. Circuits Syst. Video Technol.* **2021**. [CrossRef]
35. Kumar, A.; Bhavsar, A.; Verma, R. Detecting deepfakes with metric learning. In Proceedings of the 2020 8th international workshop on biometrics and forensics (IWBF), Porto, Portugal, 29–30 April 2020; pp. 1–6.
36. LIY, C.M.; InIctuOculi, L. Exposing AI Created Fake Videos by Detecting Eye Blinking. In Proceedings of the 2018 IEEE InterG national Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018.
37. Dufour, N.; Gully, A. Contributing data to deepfake detection research. *Google AI Blog* **2019**, *1*, 3.
38. Fridrich, J.; Kodovsky, J. Rich models for steganalysis of digital images. *IEEE Trans. Inf. Forensics Secur.* **2012**, *7*, 868–882. [CrossRef]
39. 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, Philadelphia, PA, USA, 20–22 June 2017; pp. 159–164.
40. 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, Vigo, Spain, 20–22 June 2016; pp. 5–10.
41. Rahmouni, N.; Nozick, V.; Yamagishi, J.; Echizen, I. Distinguishing computer graphics from natural images using convolution neural networks. In Proceedings of the 2017 IEEE Workshop on Information Forensics and Security (WIFS), Rennes, France, 4–7 December 2017; pp. 1–6.