Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Detecting fake images by identifying potential texture difference[☆]

Jiachen Yang[a], Shuai Xiao[a,*], Aiyun Li[a], Guipeng Lan[a], Huihui Wang[b]

[a] *School of Electrical and Information Engineering, Tianjin University, Tianjin, PR China*
[b] *Department of Engineering Jacksonville University, Jacksonville, FL, USA*

ARTICLE INFO

Article history:
Received 30 November 2020
Received in revised form 6 April 2021
Accepted 17 June 2021
Available online 25 June 2021

Keywords:
Fake detection
Security and privacy
Digital forensic
Image saliency
Texture difference

ABSTRACT

Fake detection has become an urgent task. Generative adversarial networks (GANs) extended to deep learning has shown its extraordinary ability in the fields of image, audio, and speech. But advanced technology benefits us, it also poses a threat to us when used in Cyber Crime. The Deepfake (common name for face manipulation methods) based on GANs can realize the replacement of different faces. Due to the development of GANs, faces generated by Deepfake can already be visually real. Deepfake can purposely replace any face to a different person, so that a fabricated event may be widely spread because of the convenience of the Internet, causing serious impacts such as personal attacks and cyber crime. Based on cutting-edge research , this paper proposes a intelligence forensic method of Deepfake detection. We first discover the subtle texture differences between real and fake image in image saliency, which shows difference in the texture of faces. To amplify this difference, we exploit guided filter with saliency map as guide map to enhance the texture artifacts caused by the post-processing and display the potential features of forgery. Resnet18 classification network efficiently learns the exposed difference and finally realizes the real and fake detection of face images. We evaluate the performance of the method and experiments verify that the proposed method can achieve the state-of-the-art detection accuracy .

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

The practical application of deep learning is more and more extensive, which has played a powerful role in promoting the development of human society [1]. Among them, researches and applications in the field of image and video processing are the most mature, such as target detection and image classification [2]. Image generation is another one of the research processes. The encoder–decoder is the original image generation model [3]. GANs proposed by [4] in 2014 pushed the field of image generation to a new level. GANs with adversarial learning patterns can convert arbitrary noise into the desired image, providing the possibility of image generating to be from nothing [5]. Recently, the development of GANs is rapid [6], especially in image conversion and face generation, which provides convenience for practical applications such as games, scene modeling and film production [7]. However, because of the easy application of technology, GANs is also used as a tool that threatens the safety of others [8]. The birth of Deepfake in 2017 opened a precedent for face replacement, Fakeapp and Zao were also born for face manipulation. Can you distinguish faces shown in Fig. 1 real or fake? These software can replace any face with celebrity's faces to create various videos such as pornographic clips that do not exist and circulate on the Internet, seriously polluting the network environment and jeopardizing personal information security [9]. Therefore, it is urgent to establish an effective identification system of forged face images.

For traditional image tampering methods, such as Image Splicing, Copy-Move, Object Removal, etc, there is already a relatively complete identification system in image forensics [10]. For images forged by generative models such as encoder–decoder and GANs, [11,12] tried to use traditional digital forensics methods to detect face tampering images and achieved a little sequel. However the face image generated is becoming more and more real with the improvement of GANs. Details of facial parts such as facial features, hair and wrinkles have been difficult to distinguish from the real face [13] and traditional forensics methods are difficult to efficiently identify fake images now. So [14–16] began to develop deep learning to improve the accuracy of deep forgery detection. Since then, the method of countering deep fake with deep learning has become the mainstream of forgery identification. However, the quality of deep learning model depends on

[☆] This document is the results of the research project funded by National Natural Science Foundation of China (No. 61871283), the Foundation of Pre-Research on Equipment of China (No. 61400010304) and Major Civil-Military Integration Project in Tianjin, China (No. 18ZXJMTG00170).

[*] Corresponding author.
E-mail addresses: yangjiachen@tju.edu.cn (J. Yang), xs611@tju.edu.cn (S. Xiao), liaiyun@tju.edu.cn (A. Li), lgp@tju.edu.cn (G. Lan), h.wang@ieee.org (H. Wang).

https://doi.org/10.1016/j.future.2021.06.043
0167-739X/© 2021 Elsevier B.V. All rights reserved.

Fig. 1. Actually, all the six pictures are examples of face tampering.

the size of the dataset [17]. Previous works lack large-scale and high-quality deep forgery datasets [18] and experiment results were uneven. In order to standardize the research dataset, the University of Science and Technology of China established the Celeb-DF dataset [19], Google separately established two parts of the Faceforensics dataset [20], and the DFDC preview dataset prepared by Facebook in preparation for the Deepfake Detection Challenge [21]. Nevertheless, because of the different GANs used in deep forgery methods, it is difficult to have a common and easily found forgery trace. The task of deep forgery identification is still in the development stage, and more researches is needed to find a common method to prevent deep forgery.

During the face manipulation process by deep forgery methods, the operations from taking points in the latent space to up-sample will cause artifacts or artifact patterns in the output image. Zhang et al. [22] distinguish the generated image from the real image based on the checkerboard effect contained in the image generated by GANs. But due to the different GANs used, the location and degree of artifacts are also different. In addition, due to the innovation of the algorithm, there is no visible artifact pattern caused by the GANs method. In this paper, inspired by related work [20,23], combined with the method of image quality evaluation, a deep forgery detection pipeline is proposed. Image saliency is an image partitioning mode commonly used in image quality evaluation [24], the goal is to change the real face image and the fake face image to a more easily analyzed style [25]. In order to discover the potential differences in the depth and pixel of face texture between the two kinds of images, we utilize image saliency to perform a statistical comparison analysis on a large number of real and fake face images. However, because of the insignificance of this difference, it cannot be quickly resolved and utilized by the classifier. Therefore, with the help of the detail enhancement feature of guided filter, texture enhancement is performed on both real and fake images at the same time, and the potential features of the fake face part are displayed to magnify this difference. Meanwhile, we choose Resnet18 as the backbone to classify the real and fake image, who can fully learn the enlarged texture difference between the two kinds of images, and realize the accurate classification of real and fake face images. Experimental results show that our method achieves the state-of-the-art accuracy under the same dataset compared with other works. The following is a summary of the contributions in this paper:

1. For the first time, the image saliency method is applied to discover the difference in texture depth and pixel between real and fake face images.
2. Using the detail enhancement feature of the guided filter, the texture details of fake images are enhanced, and the texture difference between the two kinds of images is enlarged.
3. By a large number of experiments, the proposed method has reached the state-of-the-art level in classification of real and fake face images.

In the rest of this paper, we introduce the related work of face tampering methods and face forgery detection methods in Section 2. In Section 3, we explain our framework and describe the application of image saliency method and texture enhancement module. The experimental results of Deepfake detection are displayed in Section 4 and compared with the other works. The last, we conclude the experimental results and discuss the shortcomings and future research directions.

2. Related work

2.1. Face tampering methods

Deepfake [26] is a deep face replacement technology based on artificial intelligence, which can change one face to another face. At first it was popular in the Reddit community. Because of its simple operation, it does not need to know a lot of knowledge to realize the replacement of faces. The method is based on two encoders with a shared decoder, which are trained to generate training images of the source faces and the target faces respectively. Faces in target sequences are replaced by faces of source images collection. Post processing will crop and align the face into the image. To create a realistic fake image, the trained encoder and decoder of source faces are applied to target faces. The decoder output is then blended with the rest of the image using Poisson Image Editing (PIE) [27].

Face2Face [28] is a facial expression manipulation based on deep learning, which keeps the expression of the source face and the expression of the target face consistent, which does not involve the replacement of faces. The implementation is based on two video input frames with manual key frame selection. These frames are made use of generating the face which can be used to re-synthesize the face under different expressions. Based on this identity reconstruction, face2face tracks the whole video to compute per frame the expression, lighting, and pose parameters as done in the original implementation.

FaceSwap [29] is a GAN-based approach to transfer the face region from a source image to a target image. Though Variational Auto-Encoder (VAE) [30] is a powerful tool to generate images, produced images exhibit blur, which allows to easily detect them as generated. Adding adversarial mechanism to VAE can effectively improve the generated image blurring, video jitter and other problems. Karras et al. [6] made a breakthrough in resolution by demonstrating the generation of high-resolution images in ProGAN. The face images are generated by improving the generator and discriminator of the model. Isola et al. [31] proposed a general-purpose solution for image-to-image translation and Wang et al. [32] improved the method by incorporating multi-scale generators and discriminator. Zhu et al. [33] extended FaceSwap to video-to-video translation problems.

2.2. Forensics methods

Because of the convenience of faking technology, anyone may face the threat of personal safety, the research on the detection of deep forgery is also keeping pace. From steganography and statistical analysis on image forensics to deep learning detection methods, researchers have increasingly improved the detection effect of face tampering.

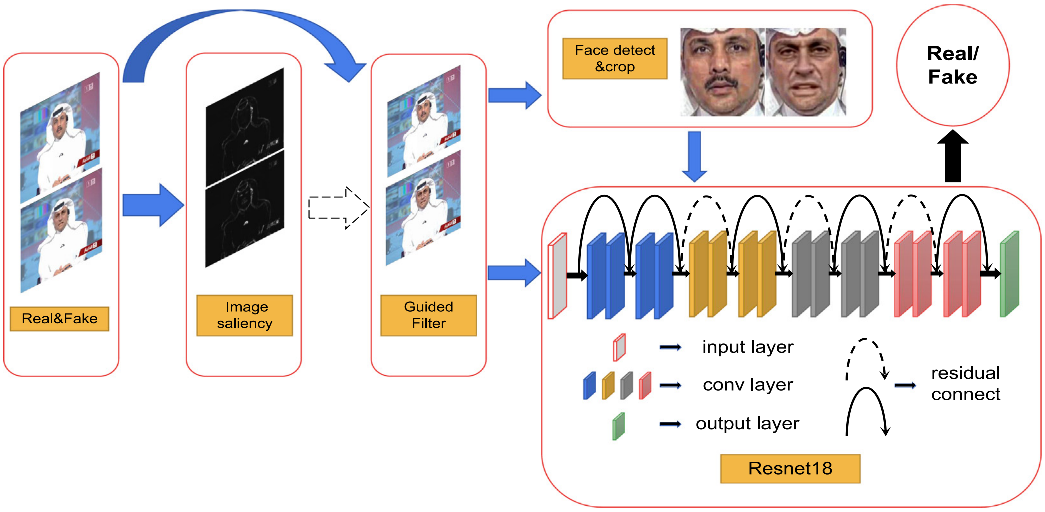


Fig. 2. This is the whole framework of the proposed method. The overall process is divided into two branches, one of which is the processing flow of the full image, which is directly input to Resnet18 after the guided filter processing, and the other is the processing flow of the face image, which is input to Resnet18 after face detecting and cropping processing. Finally, determining whether the image is real or fake.

Traditional Image forensics detection: Modern learning-based forgery detection is directly or indirectly completed on the premise of traditional forensics methods, so here we list the related development of traditional image forensics. In traditional image forensics methods, steganalysis and artifact statistics are mainly used to detect forged images. An overview of these methods can be found in recent reviews [34] [35]. The identification method based on physical characteristics is the consistency in illumination [36] or reflection [37,38]. [39–41] visualized the artifacts formed on the image residuals with the statistical characteristics of the image to detect tampered areas. Zhang et al. [18] learned specific tampering trajectories, according to verification of prior noise based on metadata, such as face area reconstruction coloring [42] or multiple compression [43]. Fridrich et al. [39] used a hand-made function to scan features with a pixel radius of 2 along the horizontal and vertical directions of the image using high-pass filter, and used these features to train a linear Support Vector Machine (SVM) classifier, which won the first IEEE Image Forensic Challenge [44]. Li et al. [45] found that human eyes of GAN-based images did not blink. Liu et al. [46] calculated a large number of deep forged images and found that the symmetry of the face is biased, the gender is difficult to distinguish and the hair is messy and disordered. However, with the improvement of GANs, these problems were gradually solved, and the application of traditional methods in deep forgery detection becomes more and more difficult.

Learning-based detection: Based on the success of deep learning in target detection and image classification. In recent years, researchers have begun to exploit Convolutional Neural Network (CNN) to combat deep forgery [47]. A lot of remarkable achievements have been made in deep forgery detection by means of intelligent confrontation. Matern et al. [48] considered the face landmark as the basis to respectively analyze the eye color, eyeball contour and tooth contour of the Deepfake face, analyze the nose and face contour of the Face2Face faces, which were regarded as features to train simple Multi-Layer Perceptron(MLP) and logistic regression classifications. Amerini et al. [49] converted the optical flow between frames into a 3 channels image with a fixed color coding method, and train the CNN to realize real and fake classification. Hsu et al. [50] compared pairs of information inputs and tried to learn the common features between faked faces by different GANs methods. Agarwal et al. [51] demonstrated the relationship between the difficulty of forgery identification and the accuracy of the GANs used in a purely theoretical way, the same time discussed the difficulty of identification in the case of Neyman–Pearson and Bayesian with a quantitative evaluation of an error boundary. Wang et al. [52] integrated a dozen different GAN-generated images as the test set, and uses a large number of images generated by one of GANs as the training set to train the classifier, to achieve the effect of learning the common features extracted by CNN structure. Jeon et al. [53] proposed a lightweight network including an image-based self-attention module to search for new feature space to detect fake images. Kumar et al. [54] divided the face part of the face2face images into four blocks, and input them into the parallel Resnet18 network. Finally they integrated them into a score to identify the fake image. [55] created a new deepfake detection network MesoNet. When publishing the Faceforensics++ dataset, Roessler et al. [20] compared the performance of multiple detection methods on the dataset, and the accuracy of Xception [56] reached the state-of-the-art. Li et al. [57] proposed a deep forgery boundary detection method, using the training data of one forgery method to achieve the state-of-the-art accuracy of detecting multiple forgery methods now.

3. Model framework

The overall framework of the proposed method in this paper is shown in Fig. 2. The main idea of the paper is the combination of image processing and network training. The guided filter is the image processing part of the framework, which magnifies the subtle texture differences on the face tampering image to form guided features that can distinguish the real image from the fake image in network. All the real and fake images preprocessed by guided filter are jointly used as the input of the network. After a amount of iterative training, the final softmax layer of the network judges whether it is real or fake. The whole process is divided into two branches, one of which is full image training, and the other branch is only training face images, which will be detected and cropped face before inputting the Resnet18. The network parameters of the two branches are not shared, and the training results do not interfere with each other, but can be used as important reference to evaluate the performance of the method.

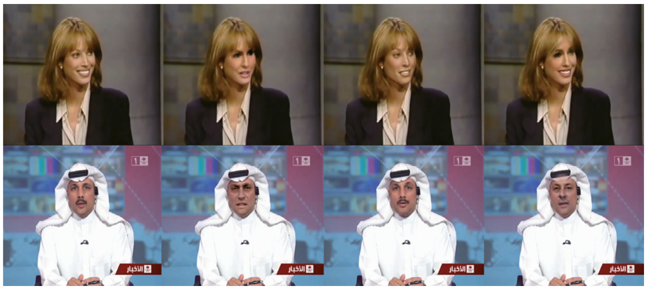


Fig. 3. The popular face tampering methods are displayed. The first column is the pristine images. From the second column to the last column: Deepfake, Face2Face, FaceSwap.

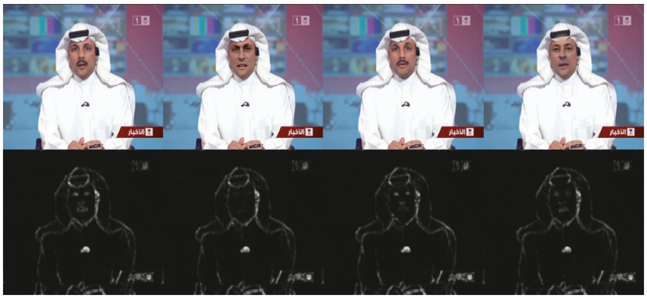


Fig. 4. The saliency maps of the real and face tampering images are displayed. From left column to right column: Pristine, Deepfake, Face2Face, FaceSwap.

3.1. Image saliency

Inspired by these works [23,58–60], we consider the image saliency method to detect possible detailed differences in forged images. The essence of saliency detection is a visual attention model, which uses the visual attention mechanism to get the most noticeable part of the image and uses a grayscale image to represent its saliency. Saliency detection is usually used as a preprocessing step to reduce the computational complexity. The input image is segmented to obtain pixels, and the contrast between each pixel and the edge pixel is calculated by multi-features, and then a saliency map based on the background is obtained. Then, the image is segmented by an adaptive threshold to select the foreground seed. The foreground-based saliency map is obtained by calculating the similarity between the foreground seed and the foreground–background saliency map. After fusing the obtained saliency maps based on the foreground and background, the Gaussian filter is used for optimization to obtain local feature saliency maps. After that, a strong classifier is obtained by training and learning the input image, and the global feature saliency map is obtained by using it. Finally, the local feature saliency map obtained in the first step is fused with the global feature saliency map obtained in the second step after training and learning, thereby obtaining the final saliency map.

In order to better grasp the texture structural information of the input image, we use super pixels instead of pixels as the minimum processing unit, and use the Simple linear iterative clustering (SLIC) algorithm for super pixel segmentation. Based on the edge priority, it is assumed that the edge area of the image is more likely to be the background. Calculating the Euclidean distance corresponding to different regions through the three characteristics of RGB, CIELab and LBP, and obtaining the significance value of each region. Dividing the input picture into K super pixels, then for the region r_i and the background b_j, the saliency value of the region can be obtained:

$$M_0 = W\left(r_i\right) \times \sum_{F}\left[\frac{1}{K}\sum_{j=1}^{K} d_F\left(r_i, b_j\right)\right] \tag{1}$$

where $W(r_i)$ is the center prior weight calculated by the normalized space distance between the center r_i and the image center. F is one of RGB, CIELAB and LBP features, and $d_F(r_i, b_j)$ is the Euclidean distance between image super-pixel area r_i and edge area b_j. The saliency value of each super pixel can be obtained by formula (1), and the pixel in the super image region can be set to this value to obtain the saliency value $M_0(x, y)$ of the pixel level, in which the coordinate (x, y) representing the pixel is the saliency map S_0. We take a large number of real face and three kinds of manipulated methods images as input respectively, and obtain the saliency maps of each image. As shown in Fig. 4, compared with the saliency map of the real face image, the deepfake saliency map only contains the outline of the face and the facial features disappear. Face2Face has facial features but is incomplete. The facial features of FaceSwap are distorted. It is known that during the process of generating and replacing human faces, texture artifacts difficult to find visually can be caused. The saliency map notices this artifact.

3.2. Texture enhancement

Inspired by the difference between real and fake face saliency map, we use a method that can enhance the details of images to magnify the difference between them. The traditional image enhancement algorithms are mainly divided into two types: frequency domain and spatial domain. The spatial domain enhancement algorithms directly process the image, such as histogram equalization and grayscale transformation [61]. The frequency domain enhancement algorithm mainly operates on images in a certain variation domain of the image, such as Fourier transform and Wavelet transform [62]. An edge-preserving filtering algorithm called guided filter has been widely used in recent years [63]. The processing requires a guided image, which is the each channel input image in this paper, information in the guided image and the kernel function can be added in the filtering process. By the guided filter, the time complexity is independent of the window size and the image is no longer independent. It is possible to use a large window to process pictures, achieve higher efficiency, and overcome the shortcomings of high time complexity of bilateral filters.

The guided filter we quoted is based on a local linear model. Assuming that a point on a function is linear with its neighboring points, a complex function can be represented by many local linear functions. When a value at a point on the function is required, simply calculate and average the values of all linear functions containing the point. Supposing the output of the guided filter function satisfies a linear relationship with the input in a two-dimensional window:

$$q_i = a_k I_i + b_k, \forall i \in w_k \tag{2}$$

where q represents the output pixel value, I represents the pixel value of the guided saliency image, which is a local window with k as the center and r as the radius, and the coefficient is the linear function. Gradient on both sides of the formula (2):

$$\nabla q = a \nabla I \tag{3}$$

The formula (3) shows guided filter with edge retention because of similar gradients in input and output image. We need to get the minimum difference between the enhanced fake image and the original fake image, that is to find the optimal solution of linear regression.

$$E(a_k, b_k) = \sum_{i\in w_k} ((a_k I_i + b_k - p_i)^2 + \varepsilon a_k^2) \tag{4}$$

$$a_k = \frac{\frac{1}{\lfloor w \rfloor}\sum_{i\in w_k} I_i p_i - \mu_k p^-}{\sigma_k^2 + \varepsilon} \tag{5}$$

$$b_k = {p_k}^- - a_k \mu_k \tag{6}$$

The optimal solution is obtained by the least squares method, where ε is the adjustment factor, μ_k is the mean value of the input fake image I in the window, σ_k^2 is the variance of the input I in the window w_k, $\lfloor w \rfloor$ is the number of pixels of the window, ${p_k}^-$ is the mean value of the original image in the window. We use this algorithm to enhance the texture characteristics of the fake image. The enhanced image in Fig. 5 has a clear texture edge compared to the original fake image. The contrast structure of the image after the texture enhancement of the real image and the fake image is more tortuous, and you can even see the obvious fake image change face frame.

3.3. Network training

According to the enhanced texture in the fake image shown in Fig. 5, we can see that although the texture in the two has been enhanced, it is still not easy to be found visually. The deep learning CNN can deepen the number of network layers, extract the features of the image layer by layer and perform linear and nonlinear calculations on these features to achieve a comprehensive analysis of the image. However, the deepening of the deep neural network will cause gradient disappearance and gradient explosion during the training process, affecting convergence and other problems. [64] Inspired by Vector of Aggregate Locally Descriptor (VLAD) and highway network, a deep residual learning framework is proposed to optimize the network training process, while deepening the number of network layers helps improve the accuracy of classification or recognition tasks. We choose Resnet18 as the backbone network to detect the real and fake images finally, 18 indicates the number of network layers containing parameters, and consists of several residual blocks. After the guided filter process, the texture features of the face part are enhanced, so we only need a small number layers of network instead of very deep networks. During network training, epoch is set to 50, which means that all training images will be iterated 50 times. After each iteration, the validation set evaluates the learning effect of the model, thereby modifying the model parameters. Considering the memory load of GPU, we set batch size to 64. In order to make the network model optimal fit, we set the learning rate to 0.0002. The network layer optimizer chooses Adam [65], and the relevant formula is as follows:

$$Vdm = \frac{\mu Vdm + (1-\mu)\, dm}{1-\mu^t} \tag{7}$$

$$Sdm = \frac{\nu Sdm + (1-\nu)\, dm}{1-\nu^t} \tag{8}$$

$$W = W - \alpha \frac{Vdm}{\sqrt{Sdm} + \varepsilon} \tag{9}$$

where Vdm is the first-order moment estimation and Sdm is the second-order moment estimation with deviation correction. μ and ν are the exponential decay rates respectively. In the experiment, we set them (0.9, 0.999). W is the updated parameter, and dm is the gradient. ε stands for the optimizer learning rate. In order to prevent over fitting, we set the dropout to 0.5 at each layer. For the loss function, we choose the cross entropy function, which is often used as the loss of the binary classification task Functions. Related formulas are as follows:

$$L(\text{class}, \text{label}) = -\text{class}[\text{label}] + \log\left(\sum_{j=0}^{n} e^{\text{input}[j]}\right) \tag{10}$$

where class and label are the input image and the label of the image respectively.

Table 1
The data in the table represents the number of images that each method participates in training, validation and testing. In each training and verification process, each forged image and real image will be mixed. The overall test and the separate test will be conducted according to the situation during performance evaluation.

Methods	Train	Validation	Test
Deepfake	194 400	14 000	14 000
Face2Face	194 400	14 000	14 000
FaceSwap	194 400	14 000	14 000
Real	194 400	14 000	14 000

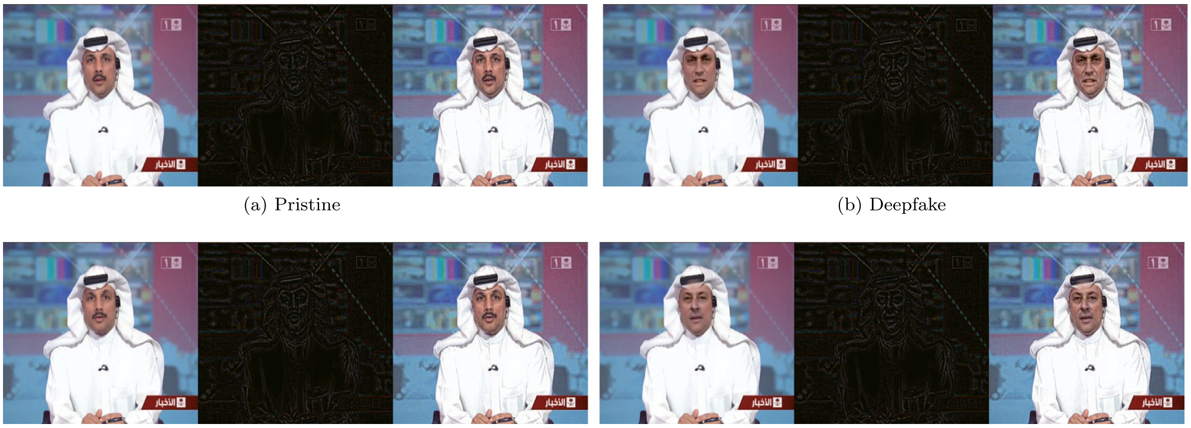
When passing through each network layer in Resnet18, each residual block will down-sample and re-sample the features of the image, and gradually learn the guided features contained in fake images. The final fully connected layer will get the score to judge the real and fake. Although Resnet18 is a shallow network in the deep learning, it can comprehensively analyze the potential features contained in the image, which can not only realize the classification of real and fake images, but also reduce the time of network training. The experimental results in the following part show that, through the processing of our method, the detection accuracy of real and fake image has reached the state-of-the-art.

4. Experimental results

In this section, we introduce the Faceforensics++ set used for all experiments in this paper, the detailed training and evaluating settings and compare the experimental results of other works.

4.1. Dataset

At present, the scale and quality of datasets applied for deep forgery detection is uneven. Many papers have proposed datasets made by GANs for their own research. There is no official general dataset like ImageNet. Among these datasets, the relatively large and high-quality dataset is Faceforensics++[20]. The DFDC dataset made by Facebook for the DeepFake Detection Challenge has not yet been fully released, so we chose Faceforensics++ dataset as the verification dataset of the proposed method. The real face video in Faceforensics++ comes from 1000 Youtube videos. The faces in these videos have a common feature. There is only one positive face in each frame of the video. Based on the faces of these 1000 videos, Deepfake, Face2Face, and FaceSwap were used for face manipulation to obtain 1000 Deepfake videos of face replacements, Face2Face videos of facial expression replacements, and GANs modified face replacement videos respectively. With the ffmpeg video frame extraction method, the three kinds of tampering videos and the real videos each contain almost 50 thousand images. Fig. 3 shows examples of three face tampering methods. In addition, in this paper all the face tampering images and real images with compress-23 as the compression way are selected to evaluate the performance of the method.



(a) Pristine

(b) Deepfake

(c) Face2Face

(d) FaceSwap

Fig. 5. These four rows of pictures respectively represent the texture enhancement process of three different face tampering methods.

4.2. Training detail

The hardware conditions of the experiments in the paper are Intel (R) Xeon (R) CPU E5 − 2620 V4 and two NVIDIA GTX Titan XP GPUs. The saliency map analysis and guided filter processing of the image are mainly based on Matlab2015, and the network algorithm is based on the Pytorch deep learning platform. Before the network training with Resnet18, the video set is randomly divided into the training set: validation set: test set = 8 : 1 : 1. Considering that the number of corresponding images in each 1000 videos is different, we have the same number of images as each video taken, 270 images per video in the training set, 100 images per video in the verification and testing set [20], and the total number is shown in Table 1. When the training set and validation set images are input to the network, they will be resized 299 × 299 quantity. Resnet18 will perform to learn and model will update and save through the hyperparameters setting. In order to facilitate performance evaluation, we train the network and statistical indicator for the full image and the face image respectively.

For the performance evaluation in this paper, we use the accuracy rate Acc to evaluate the performance of real and fake face detection. For the model trained under each tampering method, we adopt the test set for performance testing, and the *Acc* accuracy formula is as follows:

$$Acc = \frac{test_r}{test_{all}} \tag{11}$$

Where $test_r$ indicates the correct number of detection, and $test_{all}$ indicates the total number of test set.

4.3. Performance evaluation

Before starting the overall evaluation of our method, we first evaluated the feasibility of the proposed method on the Xception model trained by the low-quality face tampering dataset by [20]. We randomly took 14 000 images from 140 videos as the test set, including 2800 real images and 11 200 fake images, conducted guided filter on them and tested the accuracy in the model provided by the author of Faceforensics++. The results are shown in Table 2. According to the results in Table 2, we can see that for the trained Xception model, the accuracy of the image after the guided filter process has been greatly improved in real images, while [20] and other works have been the detection rate of the real face is around 0.5, even the highest Mesonet is only 0.7. The accuracy of the image processed by our method has reached 0.8. However, for the face tampering images with three tampering methods, the accuracy of the processed images is lower than them, where the ratio is within 15%. Therefore, we evaluate that the way we process images to improve the accuracy, especially the real face image, will be of grate helpful. Now we need to train our model with the guided filter images.

Table 2
The image precessed by guided filter are directly test on the same benchmark. Compared with other works, the test accuracy of Deepfake (DF), Face2Face (F2F), FaceSwap (FS) and real images has achieved certain results, especially in the detection of real images.

Methods	DF	F2F	FS	Real
Cozzolino [66]	0.8545	0.6788	0.7379	0.34
MesoNet [55]	0.8727	0.5620	0.6117	0.7260
Xcept.Full Image [20]	0.7455	0.7591	0.7087	0.51
XceptionNet [20]	0.9636	0.8686	0.9029	0.524
Guided image	0.8127	0.6842	0.7458	**0.8016**

Table 3
All the manipulation methods images and real images are combined as the training set. Making analysis of the influence of guided features on all the manipulation methods.

Methods	DF	F2F	FS	Real
Xcept.Full Image [20]	0.88	0.8498	0.8223	0.6585
Res18.Full Image	0.8978	0.8177	0.7890	0.8237
Res18.Face Image	0.9786	**0.9812**	0.9412	0.9348
Proposed.Full Image	**0.9843**	0.9751	**0.9535**	**0.9572**

On the premise of verifying the feasibility of our method, we combined the real 1000 videos and three kinds of face tampering videos, and trained the Resnet18 according to the data distribution in Table 1. After the network training is completed, we compared the performance of the model with and without the guided feature during the Resnet18 training process. The test results are shown in Table 3. Because of the texture difference of training sample, the network can learn the unique tampering texture of face manipulation images easily. So in each tampering mode, the test accuracy of the training model with guided features is higher than the training model without that. In addition, it can be seen from the data in Table 3 that the performance of the guided full image training far exceeds the Xception. The detection accuracies of the full image training of the Xception are all below 0.9. The detection accuracy of our method is above 0.95, and the accuracy is improved by nearly 10%. The model performance is even closer to the network trained by face images.

Table 4
Comparative analysis of detection performance with the state-of-the-art methods. Each detection method in the table is trained and tested under the condition of face images.

Methods	DF	F2F	FS
Cozzolino [66]	0.8178	0.8532	0.8569
MesoNet [55]	0.9526	0.9584	0.9343
XceptionNet [20]	0.9885	0.9836	0.9823
Face X-ray [57]	0.9912	0.9931	0.9909
Proposed method	**0.9926**	**0.9990**	**0.9937**

At the same time, we have compared with some of the frontier work of deep forgery detection at present. In order to unify the calculation standard of accuracy with these works, the calculation of accuracy for each tampering method is different from the above. The detection accuracy is calculated for the real face image and the tampered face image. The detailed data is shown in Table 4. The detection accuracy obtained by our method is basically the state-of-the-art among all methods. When Faceforensics++ dataset was published, the XceptionNet achieved good detection results. In 2020, the Face X-ray method proposed by Microsoft and Peking University detected face frames surpassed the Xception on this dataset [57]. In the performance evaluation of face image training, our method surpassed the Face X-ray method, which is the state-of-the-art, especially Face2Face manipulation methods, the detection accuracy is 0.9990, almost 100% detection accuracy. After removing background information irrelevant to the face, the real and fake face images containing the guided features can be more accurately classified by the network.

5. Conclusion

In this paper, we find the subtle texture differences that exist between the real and fake face image and manifest them through the image saliency method. Based on this finding, we employ the improved guided filter to perform image preprocessing on all fake images and real images, the purpose is to enhance the texture artifacts contained in the face manipulation image, which we call guided features. Although most of this artifact cannot be captured visually, through Resnet18 network that can continuously downsample and resample, these enlarged texture differences will be learned to realize accurate detection of real face images and fake face images. Experiments prove that our method can achieve the highest detection accuracy both in full image training and face image training, which reaches the state-of-the-art in current research. At the same time, because of the enlarged texture features, the network can quickly and accurately capture the differences and achieve convergence at a faster speed, which can reduce the time of training. However, there are still some problems in our method, which is also a common problem in the research of real and fake face detection. Network model training still requires large-scale data to achieve good results. For the unknown face tampering method, it is still necessary to train a new authentication network instead of forming a universal detection network. In future research work, we will commit to solve these problems and promote the development of face manipulation detection.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436.

[2] E.A.A. Vega, E.G. Fernández, A.L.S. Orozco, L.J.G. a Villalba, Image tampering detection by estimating interpolation patterns, Future Gener. Comput. Syst. 107 (2020) 229–237, http://dx.doi.org/10.1016/j.future.2020.01.016.

[3] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495, http://dx.doi.org/10.1109/TPAMI.2016.2644615.

[4] I. Goodfellow, J. Pougetabadie, M. Mirza, B. Xu, D. Wardefarley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, 2014, pp. 2672–2680.

[5] J.E. Tapia, C. Arellano, Soft-biometrics encoding conditional GAN for synthesis of NIR periocular images, Future Gener. Comput. Syst. 97 (2019) 503–511, http://dx.doi.org/10.1016/j.future.2019.03.023.

[6] V. Mothukuri, R. M.Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanh, G. Srivastava, A survey on security and privacy of federated learning, Future Gener. Comput. Syst. 115 (2021) 619–640, http://dx.doi.org/10.1016/j.future.2020.10.007.

[7] A.L.S. Orozco, C.Q. Huamán, D.P. Álvarez, L.J.G. Villalba, A machine learning forensics technique to detect post-processing in digital videos, Future Gener. Comput. Syst. 111 (2020) 199–212, http://dx.doi.org/10.1016/j.future.2020.04.041.

[8] A. Khodabakhsh, R. Ramachandra, K.B. Raja, P. Wasnik, C. Busch, Fake face detection methods: Can they be generalized? 2018, pp. 1–6.

[9] Y. Guo, L. Jiao, S. Wang, S. Wang, F. Liu, Fuzzy sparse autoencoder framework for single image per person face recognition, IEEE Trans. Syst. Man Cybern. 48 (8) (2018) 2402–2415, http://dx.doi.org/10.1109/TCYB.2017.2739338.

[10] J. Li, X. Li, B. Yang, X. Sun, Segmentation-based image copy-move forgery detection scheme, IEEE Trans. Inf. Forensics Secur. 10 (3) (2015) 507–518, http://dx.doi.org/10.1109/TIFS.2014.2381872.

[11] S. Marcel, J. Galbally, S. Marcel, Face anti-spoofing based on general image quality assessment, in: IEEE Intl. Conf. on Pattern Recognition, ICPR, 2014.

[12] D. Wen, H. Han, A.K. Jain, Face spoof detection with image distortion analysis, IEEE Trans. Inf. Forensics Secur. 10 (2015) 619–640, http://dx.doi.org/10.1109/TIFS.2015.2400395.

[13] Y. Zhang, L. Zheng, V.L.L. Thing, Automated face swapping and its detection, in: 2017 2nd International Conference on Signal and Image Processing, 2017, pp. 15–19.

[14] L. Zheng, S. Duffner, K. Idrissi, C. Garcia, A. Baskurt, Siamese multi-layer perceptrons for dimensionality reduction and face identification, Multimedia Tools Appl. (2015) http://dx.doi.org/10.1007/s11042-015-2847-3, URL https://hal.archives-ouvertes.fr/hal-01182273.

[15] Shuang, Bai, Growing random forest on deep convolutional neural networks for scene categorization, Expert Syst. Appl. (2017) http://dx.doi.org/10.1016/j.eswa.2016.10.038.

[16] F. Marra, D. Gragnaniello, D. Cozzolino, L. Verdoliva, Detection of GAN-generated fake images over social networks, in: 2018 IEEE Conference on Multimedia Information Processing and Retrieval, MIPR, 2018.

[17] J. Deng, W. Dong, R. Socher, L. Li, Kai. Li, Li. Fei-Fei, ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[18] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D.N. Metaxas, StackGAN++: Realistic image synthesis with stacked generative adversarial networks, IEEE Trans. Pattern Anal. Mach. Intell. abs/1710.10916 (2017) http://dx.doi.org/10.1109/TPAMI.2018.2856256, arXiv:1710.10916, URL http://arxiv.org/abs/1710.10916.

[19] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-DF: A new dataset for deepfake forensics, in: Cryptography and Security, 2019, arXiv.

[20] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Face-Forensics++: Learning to detect manipulated facial images, in: International Conference on Computer Vision, ICCV, 2019.

[21] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, C. Ferrer, The deepfake detection challenge (DFDC) preview dataset, 2019, arxiv preprint arxiv:1910.08854, 2019.

[22] X. Zhang, S. Karaman, S.-F. Chang, Detecting and simulating artifacts in GAN fake images, 2019.

[23] H.H. Nguyen, J. Yamagishi, Echizen, Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2019, pp. 2307–2311.

[24] J. Yang, K. Sim, X. Gao, W. Lu, Q. Meng, B. Li, A blind stereoscopic image quality evaluator with segmented stacked autoencoders considering the whole visual perception route, IEEE Trans. Image Process. 28 (3) (2019) 1314–1328, http://dx.doi.org/10.1109/TIP.2018.2878283.

[25] J. Yang, K. Sim, W. Lu, B. Jiang, Predicting stereoscopic image quality via stacked auto-encoders based on stereopsis formation, IEEE Trans. Multimed. 21 (7) (2019) 1750–1761, http://dx.doi.org/10.1109/TMM.2018.2889562.

[26] D. github, Deepfake, 2017, https://github.com/deepfakes/faceswap.

[27] P. Perez, M. Gangnet, A. Blake, Poisson Image editing, 22 (3) (2003) 313–318, http://dx.doi.org/10.1145/1201775.882269.

[28] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. Niebner, Face2Face: Real-time face capture and reenactment of RGB videos, in: IEEE Conference on Computer Vision Pattern Recognition, 2016, pp. 2387–2395.

[29] Faceswap, 2018, https://github.com/marekkowalski/faceswap/.

[30] D.P. Kingma, M. Welling, Stochastic gradient VB and the variational auto-encoder, J. Beijing Admin. College (2013).

[31] P. Isola, J. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, CVPR2107 (2017) 5967–5976, http://dx.doi.org/10.1109/CVPR.2017.632.

[32] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, B. Catanzaro, Video-to-video synthesis, in: Advances in Neural Information Processing Systems, NeurIPS, 2018.

[33] T.C. Wang, M.Y. Liu, J.Y. Zhu, A. Tao, J. Kautz, B. Catanzaro, High-resolution image synthesis and semantic manipulation with conditional GANs, CVPR2018 (2017).

[34] J. Redi, W. Taktak, J. Dugelay, Digital image forensics: a booklet for beginners, Multimedia Tools Appl. 51 (1) (2011) 133–162, http://dx.doi.org/10.1007/s11042-010-0620-1.

[35] N. Memon, Photo forensics, in: Acm International Workshop, 2015.

[36] E. Kee, H. Farid, Exposing digital forgeries from 3-D lighting environments, in: IEEE International Workshop on Information Forensics and Security, 2010, pp. 1–6, http://dx.doi.org/10.1109/WIFS.2010.5711437.

[37] M.K. Johnson, H. Farid, Exposing Digital Forgeries Through Specular Highlights on the Eye, Vol. 4567, 2007, pp. 311–325, http://dx.doi.org/10.1007/978-3-540-77370-2_21.

[38] J.F. O'Brien, H. Farid, Exposing photo manipulation with inconsistent reflections, ACM Trans. Graph. 31 (1) (2012) http://dx.doi.org/10.1145/2077341.2077345.

[39] J. Fridrich, J. Kodovsky, Rich models for steganalysis of digital images, IEEE Trans. Inf. Forensics Secur. 7 (3) (2012) 868–882, http://dx.doi.org/10.1109/TIFS.2012.2190402.

[40] D. Cozzolino, G. Poggi, L. Verdoliva, Splicebuster: A new blind image splicing detector, in: IEEE International Workshop on Information Forensics and Security, WIFS, 2015, pp. 1–6.

[41] D. Cozzolino, L. Verdoliva, Noiseprint a CNN-based camera model fingerprint, IEEE Trans. Inf. Forensics Secur. (2018) http://dx.doi.org/10.1109/TIFS.2019.2916364.

[42] M. Barni, E. Nowroozi, B. Tondi, Detection of adaptive histogram equalization robust against JPEG compression, in: 2018 International Workshop on Biometrics and Forensics, IWBF, 2018, http://dx.doi.org/10.1109/IWBF.2018.8401564.

[43] S. Mandelli, N. Bonettini, P. Bestagini, V. Lipari, S. Tubaro, Multiple jpeg compression detection through task-driven non-negative matrix factorization, in: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2018, pp. 2106–2110.

[44] D. Cozzolino, D. Gragnaniello, L. Verdoliva, Image forgery detection through residual-based local descriptors and block-matching, in: 2014 IEEE International Conference on Image Processing, ICIP 2014, 2015, pp. 5297–5301, http://dx.doi.org/10.1109/ICIP.2014.7026072.

[45] Y. Li, M. Chang, S. Lyu, In Ictu Oculi: Exposing AI generated fake face videos by detecting eye blinking, in: IEEE International Workshop on Information Forensics and Security, WIFS, 2018.

[46] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: IEEE International Conference on Computer Vision, 2015, pp. 3730–3738, http://dx.doi.org/10.1109/ICCV.2015.425.

[47] J. Yang, J. Man, M. Xi, X. Gao, W. Lu, Q. Meng, Precise measurement of position and attitude based on convolutional neural network and visual correspondence relationship, IEEE Trans. Neural Netw. Learn. Syst. (2019) 1–12, http://dx.doi.org/10.1109/TNNLS.2019.2927719.

[48] F. Matern, C. Riess, M. Stamminger, Exploiting visual artifacts to expose deepfakes and face manipulations, in: 2019 IEEE Winter Applications of Computer Vision Workshops, WACVW, 2019.

[49] I. Amerini, L. Galteri, R. Caldelli, A.D. Bimbo, Deepfake video detection through optical flow based CNN, in: IEEE/CVF International Conference on Computer Vision Workshop, ICCVW, 2019.

[50] C. Hsu, Y. Zhuang, C. Lee, Deep fake image detection based on pairwise learning, Appl. Sci. 10 (1) (2020) 370, http://dx.doi.org/10.3390/app10010370.

[51] S. Agarwal, L.R. Varshney, Limits of deepfake detection: A robust estimation viewpoint, in: Learning, 2019, arXiv.

[52] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, A.A. Efros, CNN-generated images are surprisingly easy to spot... for now, 2019.

[53] H. Jeon, Y. Bang, S. Woo, FDFtnet: Facing off fake images using fake detection fine-tuning network, in: Computer Vision and Pattern Recognition, 2020, arXiv.

[54] P. Kumar, M. Vatsa, R. Singh, Detecting face2face facial reenactment in videos, in: Computer Vision and Pattern Recognition, 2020, arXiv.

[55] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, MesoNet: a compact facial video forgery detection network, in: IEEE International Workshop on Information Forensics and Security, WIFS, 2018, pp. 1–7.

[56] F. Chollet, Xception: Deep learning with depthwise separable convolutions, 2017, pp. 1800–1807.

[57] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, B. Guo, Face X-ray for more general face forgery detection, CVPR2020 (2019).

[58] G. Huang, Z. Liu, V.D.M. Laurens, K.Q. Weinberger, Densely connected convolutional networks, 2016.

[59] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation, 2014, arXiv:1406.1078, CoRR abs/1406.1078, URL http://arxiv.org/abs/1406.1078.

[60] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, P. Natarajan, Recurrent convolutional strategies for face manipulation detection in videos, 2019.

[61] T. Celik, Spatial entropy-based global and local image contrast enhancement, IEEE Trans. Image Process. 23 (12) (2014) 5298–5308.

[62] W. Alexander, W. Jrgen, H. Ulrich, H. Michael, T. Buzug, Reconstruction enhancement by denoising the magnetic particle imaging system matrix using frequency domain filter, IEEE Trans. Magn. (2015).

[63] A. Jameel, M.M. Riaz, A. Ghafoor, Guided filter and IHS-based pan-sharpening, IEEE Sens. J. 16 (1) (2016) 192–194.

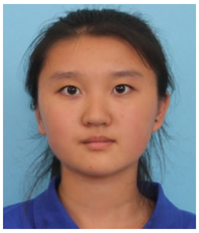
[64] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778.

[65] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, Comput. Sci. (2014).

[66] D. Cozzolino, G. Poggi, L. Verdoliva, Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection, 2017.

Jiachen Yang received the M.S. and Ph.D. degrees in communication and information engineering from Tianjin University, China, in 2005 and 2009, respectively. He is currently a professor at Tianjin University. His research interests include image quality evaluation, wireless content caching, stereo vision research, pattern recognition and computer vision.

Shuai Xiao is currently pursuing the Ph.D. degree at the School of Electrical and Information Engineering, Tianjin University, China. His research interests are Digital forensics, computer vision and pattern recognition.



Aiyun Li is currently pursuing the M.S. degree at the school of electrical and information engineering, Tianjin University, China. Her research interests include forged image forensics and computer vision.

Guipeng Lan is currently pursuing the M.S. degree at the School of Electrical and Information Engineering, Tianjin University China. His research interests are Digital forensics, computer vision and pattern recognition.

Huihui Wang received the Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, VA, USA, in 2013. Her research interests include cyber–physical systems, the Internet of Things, healthcare and medical engineering based on smart materials, robotics, haptics based on smart materials/structures, ionic polymer metallic composites, and MEMS.