Pattern Recognition 116 (2021) 107950

Contents lists available at ScienceDirect

Pattern Recognition

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

PRRNet: Pixel-Region relation network for face forgery detection

Zhihua Shang[a], Hongtao Xie[a,*], Zhengjun Zha[a], Lingyun Yu[a,b], Yan Li[a,c], Yongdong Zhang[a]

[a] School of Information Science and Technology, University of Science and Technology of China, Hefei, China
[b] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China
[c] Beijing Kuaishou Technology Co., Ltd., Beijing, China

ARTICLE INFO

Article history:
Received 1 July 2020
Revised 8 January 2021
Accepted 1 March 2021
Available online 24 March 2021

Keywords:
Face forgery detection
Forgery localization
Inconsistency detection
Relation learning

ABSTRACT

As advanced facial manipulation technologies develop rapidly, one can easily modify an image by changing the identity or the facial expression of the target person, which threatens social security. To address this problem, face forgery detection becomes an important and challenging task. In this paper, we propose a novel network, called Pixel-Region Relation Network (PRRNet), to capture pixel-wise and region-wise relations respectively for face forgery detection. The main motivation is that a facial manipulated image is composed of two parts from different sources, and the inconsistencies between the two parts is a significant kind of evidence for manipulation detection. Specifically, PRRNet contains two serial relation modules, i.e. the Pixel-Wise Relation (PR) module and the Region-Wise Relation (RR) module. For each pixel in the feature map, the PR module captures its similarities with other pixels to exploit the local relations information. Then, the PR module employs a spatial attention mechanism to represent the manipulated region and the original region separately. With the representations of the two regions, the RR module compares them with multiple metrics to measure the inconsistency between these two regions. In particular, the final predictions are obtained totally based on whether the inconsistencies exist. PRRNet achieves the state-of-the-art detection performance on three recent proposed face forgery detection datasets. Besides, our PRRNet shows the robustness when trained and tested on different image qualities.

1. Introduction

Nowadays, there raise many advanced facial manipulation approaches [1–4] to modify the visual information in facial images by synthesizing deceptive faces that are difficult to be distinguished by humans. However, when these manipulated images are used for cheating or slandering, they would cause security issues and even crisis of confidence in our society. Therefor, it is supremely important to develop effective methods for face forgery detection.

In traditional forensics techniques [5], the inconsistencies are a significant kind of evidence to detect manipulated images. In particular, the manipulated image generally contains two parts, the fake face which is the manipulated region and the real background which is the original region. As shown in Fig. 1, there are two main categories of facial manipulations, i.e. changing the identity and modifying the facial expression. To obtain the desired images, both of these two kinds of facial manipulation methods synthesize fake faces and blend them into real original images. Because the sources of the manipulated region and the rest regions are different, there is a kind of intra-image relation between these two regions, i.e. their features are inconsistent with each other, as indicated in [6]. Many early methods focus on the existence of inconsistencies and achieve image forgery detection with handcrafted features, such as in color filter array (CFA) interpolation artifacts [7], illumination [8] or local noise variances [9]. It is demonstrated that the relation between regions from different sources could be used to recognize manipulated images. However, with the rapid development of synthesis approaches, [10–12] the performances of specific handcrafted features are not satisfactory.

Recent works [13,14] introduce deep learning to image forgery detection and attempt to employ neural networks to extract suitable and meaningful features from input images. Most deep-learning-based works [13,15] detect manipulated images by directly classify the input image based on their global features as in a regular binary classification task. Besides, some works impose the local features extracted by neural networks to localize the forgery with various strategies, such as multi-task learning [16,17], or attention mechanism [18]. However, in these deep-learning-based approaches, the relation between the manipulated region and the original region mentioned before is not efficiently utilized.

In this paper, we focus on effectively exploiting the spatial relation and inconsistencies for face forgery detection. We propose a novel network, called Pixel-Region Relation Network (PRRNet), as

[*] Corresponding author.
E-mail address: htxie@ustc.edu.cn (H. Xie).

https://doi.org/10.1016/j.patcog.2021.107950
0031-3203/© 2021 Elsevier Ltd. All rights reserved.

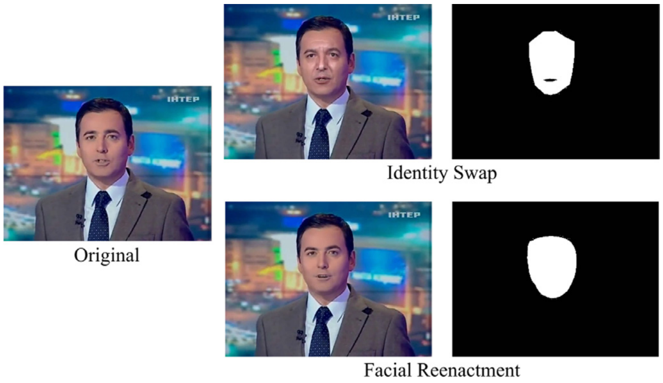

Fig. 1. The left is the original image. And the middle column shows the manipulated images respectively generated by identify swap(top) and facial reenactment(bottom), with their groundtruth of manipulated regions in the right.

illustrated in Fig. 2. It captures the intra-image relation in two levels, i.e. the pixel-wise relation between different pixels in feature map, and the region-wise relation between the manipulated region and the original region, for localizing the manipulated regions and detecting the forgeries. Specifically, we append two serial relation modules on top of our backbone network. The first one is a Pixel-wise Relation (PR) module and the second one is a Region-wise Relation (RR) module.

The PR module is proposed to capture the pixel-wise relation to exploit contextual information and increase the discriminant ability of features. In particular, it encodes the similarities between every two pixels in the feature map to obtain the representations of spatial dependencies called relation feature. Then, the PR module aggregates the features with a spatial attention mechanism to represent the manipulated region and the original region respectively. Next, we feed the features of the two regions to the RR module in pairs. The RR module detects the inconsistencies by comparing the features of the manipulated region and the original region. In particular, the RR module fuses three common linear metrics to capture the relation between the two regions from multiple aspects. Finally, the decision on whether the input is manipulated depends entirely on the inconsistencies.

We evaluate our PRRNet on three recent proposed large-scale datasets, i.e. FaceForensics++ [13], Celeb-DF [14], and DFDC [19], and achieve the state-of-the-art performance. In particular, PRRNet achieves 86.13% accuracy on the low-quality images of FaceForensics++. In addition, extensive experiments show that PRRNet is robust when trained and tested in different image quality levels, which benefits from the intra-image relation. Because the change of quality is less obviously concerned with the intra-image relation, the relation could be kept approximately.

Our main contributions can be summarized as follows:

- We propose a novel Pixel-Region Relation Network (PRRNet) to exploit intra-image relations in pixel level and region level for localizing and classifying face forgeries.
- A pixel-wise relation module is proposed to represent the relation between every two pixels to enhance the discriminant ability of local features. It also employs an attention mechanism to extract features for different regions.
- A region-wise relation module is proposed to measure the inconsistency between regions by fusing multiple metrics.
- We achieve new state-of-the-art results on three datasets, especially reach 86.13% accuracy on low-quality images of FaceForensics++. And our experiments demonstrate the robustness of our approach to different image qualities.

2. Related work

2.1. Inconsistency detection

The local inconsistency in the manipulated image is an interpretable and significant kind of evidence for image forgery detection, which is widely used in early digital forensics methods. Many works [5,7,8,20] impose on a certain underlying statistic of input image to detect manipulated images. Popescu and Farid [7] pointed out that most digital cameras would introduce the specific correlations to three-channel color images by the color filter array (CFA) interpolation. Hence they revealed traces of digital tampering by detecting these correlations. Furthermore, Ferrara et al. [21] employed the CFA artifacts in localizing the image forgery. Lukas et al. [22] indicated that each digital camera has its unique identification fingerprint which is based on the sensor's pattern noise and can be used to identify the image source. Besides these noise-level inconsistencies, Johnson and Farid [8] focused on the visual clues, the inconsistencies in lighting, and exploit them as an available tool for detecting digital tampering. Yang et al. [23] revealed the manipulation by estimating the 3D head poses from the face images. These methods are based on explicit observations about the characteristics and defects of manipulated images, hence they have a high interpretability. However, the inconsistencies of manipulated images are various and complex, while these methods generally focus on specific kind of inconsistency, which can not represent the differences between the forgeries and the original images accurately. In contrast, we attempt to learn complex inconsistencies based on deep learning rather than specific handcraft features.

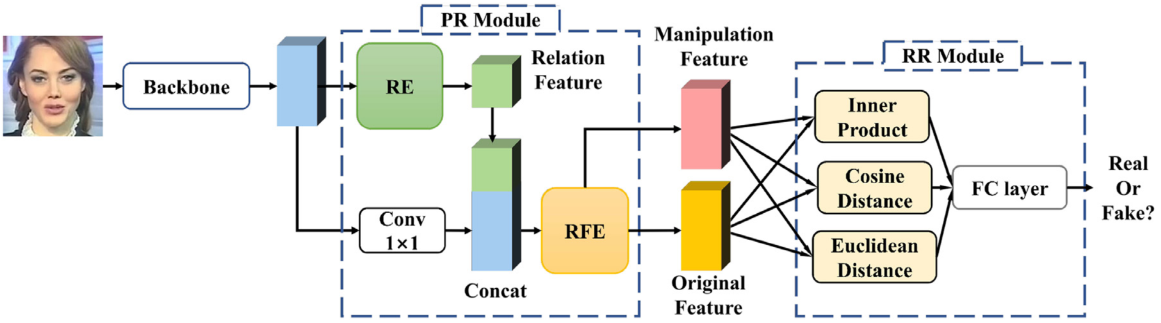

Fig. 2. An overview of the Pixel-Region Relation Network (PRRNet). Given an input facial image, we first extract its feature map with a backbone network. Next, a pixel-wise relation (PR) module is applied to calculate the relation feature and combine it with the visual feature. Then it aggregates all features with a spatial attention mechanism for the manipulated region and the original region respectively. In the end, An region-wise relation (RR) module compares the two features with multiple metrics and fuses them to obtain the final decision.

2.2. Deep-Learning-Based methods for face forgery detection

Recently, there are many facial manipulation technologies achieving great success, e.g. Deepfakes [1], FaceSwap [2], Face2Face [3], and NeuralTextures [4]. The manipulated images generated by these technologies are deceptive and bring a big challenge for traditional forgery detectors. Therefore, many works [13,14,18,19] pay attention to the face forgery detection which detect face forgeries as a regular binary classification task. Some works [13,15] tries to use various networks for improvement of accuracy. These methods have similar pipelines with two steps, i.e. extracting a global feature of the input image by convolutional neural network (CNN) and differentiating this feature by a classifier. However, these methods do not pay attention to the fact that the original regions and the manipulated regions come from different sources. Furthermore, considering this difference, some works attempt to localize the forgeries. Dang et al. [18] used an attention mechanism to highlight the manipulated regions based on the possibility of manipulation. Li et al.[6] focused on the boundary of the fake faces named face X-ray for the improvement of generalizability, and determine the input image based on whether its face X-ray exists. Although the manipulated regions are localized, the relation between them and the original regions is underused. In contrast, we exploit the spatial correlation to locate the manipulated regions, and classify the input image based on the inconsistencies between the manipulated region and the original region.

2.3. Relational learning

Many previous works [24–28] indicate that modeling the relation of objects as human can improve the performances of algorithms. Recent methods [26,27,29] focus on the local relations between regions. Hu et al. [29] modeled the relation through interaction of several objects simultaneously between their appearance feature and geometry. Fu et al. [25] adaptively integrated local features with their global dependencies in spatial and channel dimensions respectively. The above methods employ the attention mechanism to model the dependency between the elements, but ignore the structural characteristics of relations. In contrast, our PR module employs the convolution layers to extract the features of local relation maps between each pixel and other pixels. The extracted features are used for subsequent pixel-wise classification and region-wise relations extraction. Besides, Sung et al. [30] decided whether a couple of images are in the same class based on the inter-image relation information. While our RR module focuses on the intra-image relation between the manipulated region and the original region to detect the inconsistencies in the fake images.

3. Methodology

In this section, we first present the preliminary knowledge about detecting manipulated facial images. Then we present the general framework of our proposed network. Next, we provide the details of the two relation modules which capture the pixel-wise and the region-wise relations respectively. Finally, we define the loss function of our model.

3.1. Preliminary knowledge

The general image manipulation methods can be divided into many types, such as copy-clone, splicing, and object removal. While in facial manipulation, there are mainly two types, i.e. splicing and entire synthesis. In practical application, the former is more frequently used. Hence in this work, we focus on recognizing the spliced facial images. The typical pipeline to manipulate face images is shown in Fig. 3. The first step is locating the face region and cropping it. The second step is using an image synthesis method to generate a fake face for changing the identity or expression. The last step is splicing the fake face into the original image. Thus, the intra-image inconsistencies exist in the spliced manipulated images, as indicated in [6].

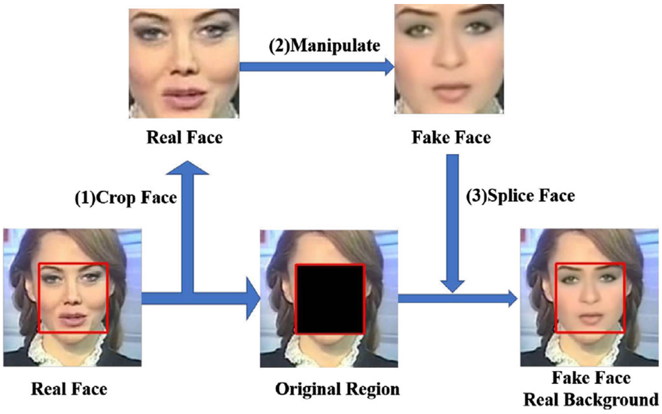

Fig. 3. Illustrating the typical pipeline to manipulate facial images.

However, previous deep-learning-based detectors regard the face forgery detection as a regular binary classification without effectively imposing the inconsistencies. Given an input image, most of previous detectors first extract the feature of input image with a CNN and then classify this feature with a fully connected layer. In this way, the models use global features to capture the input images without distinguishing the manipulated regions and the original regions.

3.2. Pixel-Region relation network

In order to exploit the spatial relation, we design two types of relation modules in different scales, i.e. Pixel-wise Relation (PR) module and Region-wise Relation (RR) module, as illustrated in Fig. 2. Given an input feature map generated by the backbone network, the PR module locates the suspected manipulated region and obtains the features of the manipulated region and the original region, as follows:

$$F_{man}, F_{ori} = f_{PR}(f_e(x)), \tag{1}$$

Where, f_{PR} represents the PR module, F_{man} and F_{ori} denote the features of the manipulated region and the original region respectively. With F_{man} and F_{ori} as input, the RR module compares these two features to capture the relation between the two regions for determining whether the input image is manipulated, as follows:

$$s = f_{RR}(F_{man}, F_{ori}), \tag{2}$$

where f_{RR} denotes the RR module and s is the inconsistency score which describes how the two regions are inconsistent. Finally, we can detect the manipulated facial images based on their inconsistency score. In this way, the relations in different levels are used for both localizing manipulated regions and classifying images.

3.3. Pixel-wise relation module

As shown in Fig. 2, the PR module consists of two parts, i.e. Relation Encoder (RE) and Region Feature Extractor (RFE), which are used to obtain relation features and region-wise features respectively. The input of the PR module is the feature map extracted by the backbone network, and its output is the features of the manipulated region and the original region.

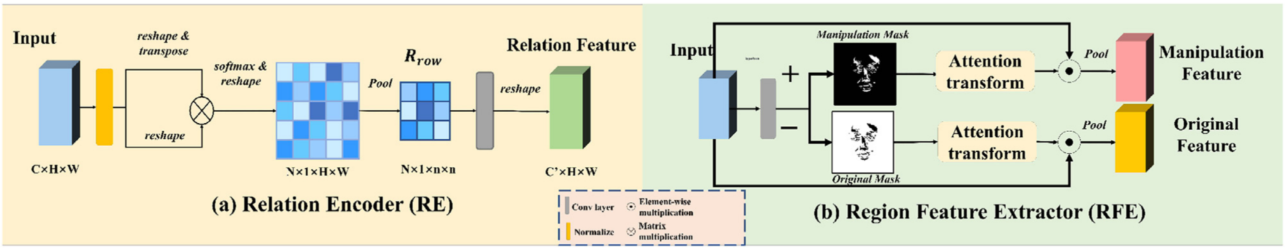

Fig. 4. The details of Relation Encoder (a) and Region Feature Extractor (b).

The local features are generated by traditional fully convolutional networks that ignore the specific spatial structure of manipulated facial images. Hence, we propose a Relation Encoder to map the spatial relation between features in every two pixels into relation features before pixel-wise classification. The structure of Relation Encoder is shown in Fig. 4 (a). Given the input feature map $F_{input} \in \mathbb{R}^{C\times H\times W}$, where C, H, and W denote the channel number, height, and width respectively, we first normalize the feature in each pixel. Then we reshape the normalized feature map, denoted as $Q \in \mathbb{R}^{C\times N}$, where $N = H \times W$ is the number for pixels in the feature map. The relation map $R \in \mathbb{R}^{N\times N}$ is calculated by performing a matrix multiplication with Q and Q^T, where the superscript T means the transpose. And R is normalized along the second dimension with softmax function. It can be formulated as:

$$R_{i,j} = \frac{exp(Q_i^T \times Q_j)}{\sum_{j=1}^{N} exp(Q_i^T \times Q_j)}, \tag{3}$$

where $Q_i \in \mathbb{R}^{C\times 1}$ represents the feature of i^{th} pixel, $R_{i,j}$ represents the relation between the i^{th} pixel and the j^{th} pixel.

To effectively exploit the relation, we encode R as a relation feature for each pixel separately. Each row of R represents a pixel-wise relation map between a certain pixel and all pixels in the feature map. Instead of directly mapping rows into feature space with fully connected layers, which would break the spatial structure, we reshape the i^{th} raw to a two-dimensional matrix $R^i_{row} \in \mathbb{R}^{H\times W}$. Further, the structures of input images are face-center which are simple and similar. Therefore R^i_{row} in lower resolution is adequate to represent the pixel-wise relation for face forgery detection, which could reduce the computational cost. We apply average-pooling on R^i_{row} to $n \times n$. Then we extract the feature of R^i_{row} with few convolutional layers to calculate the relation feature $F^i_{rel} \in \mathbb{R}^{C'\times H\times W}$ for the i^{th} pixel. Finally, we update the feature map as:

$$F_{upd} = \mathcal{C}(f_{1\times 1}(F_{input}), F_{rel}), \tag{4}$$

where $\mathcal{C}(\cdot,\cdot)$ represents the concatenation of feature maps in channel, and $f_{1\times 1}()$ is a 1×1 convolutional layer.

With the updated feature map F_{upd}, we propose a Region Feature Extractor to aggregate the features for different regions with a spatial attention mechanism. The structure of Region Feature Extractor is shown in Fig. 4 (b). We compute a manipulation mask $M_{man} \in \mathbb{R}^{C\times H\times W}$ by pixel-wise classifying F_{upd}. F_{upd} is fed to a 1×1 convolution layer and an adjacent activation layer, as follows:

$$M_{man} = Sigmoid(\phi_{1\times 1}(F_{upd})), \tag{5}$$

where $Sigmoid(\cdot)$ means the sigmoid activation layer and $\phi_{1\times 1}(\cdot)$ represent the convolution layer. Each pixel in M_{man} has a value in range [0.0,1.0], which represents the probability of the corresponding image patch being manipulated. Based on the M_{man}, the PR module can represent the manipulated region effectively.

Instead of using M_{man} as the attention map like in [18], we generate the attention map A_{man} for the manipulated region based on the M_{man} with normalization, as follows:

$$A_{man} = \frac{M_{man}}{\frac{1}{N}\sum_{i=0}^{N} M^i_{man}}. \tag{6}$$

In A_{man}, the value of each pixel is positively related to the probability of being manipulated, which makes the model focus on the informative regions. In addition, compared with M_{man}, there is a significant advantage of A_{man}. When the values of all pixels in M_{man} are close to 0, weighting feature map with M_{man} leads all features of any position to be small. While A_{man} always remains a high weight for the region which is more likely to be manipulated no matter the absolute value of the probability. In the end, we compute the feature of the manipulated region F_{man} as follows:

$$F_{man} = \frac{1}{N}\sum_{i=0}^{N} A^i_{man} \times F^i_{upd}. \tag{7}$$

In addition, the feature of the original region F_{ori} and corresponding attention map A_{ori} are obtained as a similar way where the original mask $M_{ori} = 1 - M_{man}$ is employed instead of M_{man}.

3.4. Region-wise relation module

The RR module is designed to classify the images by measuring the inconsistencies between the manipulated region and the original region. To describe how the two regions are inconsistent, we define an inconsistency score $s \in [0, 1]$, as in Eq. 2. Noted, the relation between features of different regions plays an important role in the calculation of s. In particular, when $s = 1$, it means that the two regions come from different sources and have some inconsistent features.

Given F_{man} and F_{ori}, RR module first maps them into a latent space with $f_m(\cdot)$ for reducing superfluous information. As in Eq. 7, both F_{man} and F_{ori} are weighted sum of F_{upd}, which are in the same feature space. While the feature of each pixel in F_{upd} is used to locate the manipulated region, which contains some information unrelated to the inconsistencies. Then the RR module learns a non-linear metric for the region-wise relation, i.e. the inconsistencies. Because the inconsistencies have a positive correlation with the distance between features of the two regions and a negative correlation with their similarity. Hence, to obtain a complex non-linear metric, we fuse three linear metrics, i.e. Cosine Distance, Euclidean Distance, and Inner Product, to capture the relationships between the two regions from different aspects.

More specifically, Cosine Distance is widely used to the feature similarity in the hypersphere space. Given two features $A, B \in \mathbb{R}^{(1\times n)}$, the Cosine Distance is only related to the angle between them. In contrast, even if the directions of A,B are the same, the Euclidean Distance is possibly large. While the Inner Product is related to both the directions and the lengths. Hence, these three metrics could measure the relation between features in different aspects.

To obtain a comprehensive and suitable metric for measuring inconsistencies, we fuse the results of these metrics by concatenating them to a feature and input it into a subsequent classifier. Therefore s is calculated as follows:

$$s = f_\phi(\mathcal{C}(m_{cos}, m_{euc}, m_{inn})), \tag{8}$$

where F_{man} and F_{ori} are mapped by $f_m(\cdot)$, f_ϕ is a classifier, $m_{\cos}$, m_{euc}, m_{sim} respectively denote the Cosine Distance, Euclidean Distance, and Inner Product between F_{man} and F_{ori}. For inference, when $s \geq 0.5$, the input image is classified as a manipulated image. Different from regular classifiers used for face forgery detection, the RR module detects the manipulated image paying attention to the region-wise relation between the manipulated region and the original region rather than the global feature.

3.5. Loss function

Our approach is designed to simultaneously classify the input images and locate the manipulated regions. We employ the supervised learning for training with the loss function composed of two parts. For measuring the accuracy of the manipulation mask, we use the cross entropy loss as follows:

$$L_{loc} = -\sum_{I \in D} \frac{1}{N} \sum_{i=0}^{N} \left(G^i \cdot logM^i_{man} + \left(1 - G^i\right) \cdot \log\left(1 - M^i_{man}\right)\right), \tag{9}$$

where D is the training dataset, G represents the label of manipulation region. And we also use the cross entropy loss for classification, as follows:

$$L_{cls} = -\sum_{I \in D} (y \cdot \log(s) + (1 - y) \cdot \log(1 - s)), \tag{10}$$

where y is the label whose value is 1 for the manipulated images and is 0 for the original images. In the end, A hyper-parameter λ is used to balance these two loss function, so the whole loss $L = L_{loc} + \lambda L_{cls}$. In our experiments, λ is set to 0.1. The more analyses can be found in subsection 4.3.4.

4. Experiment

In this section, we first introduce the datasets and implementation details. Then in order to demonstrate the effectiveness and robustness of PRRNet for face forgery detection, we evaluate our proposed approach, PRRNet, on three datasets, i.e. FaceForensics++ [13], Celeb-DF [14], and DFDC [19]. Finally, we analyze the effects of components and hyper-parameter in PRRNet.

4.1. Datasets and implementation details

4.1.1. Datasets

FaceForensics++ (FF++) The dataset is a recently proposed popular benchmark dataset. It collects 1000 pristine videos from the Internet and employs four automated state-of-art face manipulation technologies, i.e. Deepfakes (DF) [1], FaceSwap (FS) [2], Face2Face (F2F) [3], and NeuralTextures (NT) [4], to generate 4000 manipulated videos totally. In addition, in order to simulate realistic manipulated videos, it compresses the raw video with two different quality levels using the H.264 codec. In particular, it uses a quantization parameter equal to 23 to generate high-quality videos (HQ). And low-quality videos (LQ) are produced with a quantization parameter equal to 40.

Celeb-DF (Celeb) The dataset is also a new larger-scale dataset. There are 590 real videos and 5639 DeepFake videos, corresponding to more than 2.3 million frames in total. Especially, the DeepFake videos are generated by an improved Deepfake manipulation method. Moreover, to evaluate the robustness of our approach, we compress the videos with two quantization parameters, 20 (c20) and 40 (c40), separately.

DeepFake Detection Challenge (DFDC) There are 5214 videos in this dataset. Two unspecified methods are employed to generate manipulated videos. In order to approximate real-life video distributions, a part of the test set is augmented in three ways, i.e. (1) reducing the FPS of video to 15 (low-fps); (2) changing the resolution of the video to 1/4 of its original size (low-res); (3) compressing the video to lower quality (low-quality). Especially, each video is processed in no more one way.

Table 1
Comparison with previous state-of-the-art face forgery detectors in binary detection accuracy on the FF++. All models are trained for all four manipulation methods. † represents our re-implementation.

Methods	raw	HQ	LQ
Steg. Features + SVM [20]	97.63	70.97	55.98
Cozzolino et al. [33]	98.57	78.45	58.69
Bayar and Stamm [34]	98.74	82.97	66.84
Rahmouni et al. [35]	97.03	79.08	61.18
MesoNet [15]	95.23	83.10	70.47
Xception [13]	<u>99.26</u>	95.73	81.00
Face X-ray† [6]	99.1	95.97	82.94
Xception † [13]	99.11	**96.47**	82.69
PRRNet (ours)	**99.17**	96.15	**86.13**

It is noted that the performances of detectors are evaluated on frame-level rather than video-level on these three datasets in this work.

4.1.2. Implement details

In this work, HRNet-w30 [31] is employed as our backbone network, which generates four feature maps in different resolution levels for one input image. These feature maps are resized to 75 × 75 and are concatenated together as the input of the PR module. In the PR module, we set $n = 5$ to split the whole relation map to patches. The channel numbers of the relation feature equal to 64. In the RR module, the input features are first mapped by two fully connected layers where the former is followed by a BatchNorm layer and a ReLu activation layer.

For training, the Adam [32] is employed with the default values for the moments ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$). We train models for 7 epochs. In particular, in the first 4 epochs, the parameters of the backbone network initialized by ImageNet are fixed. The rest network is trained with a learning rate of 0.0002. Then, the whole network is finetuned with a learning rate of 0.00002.

We preprocess the images before inputting them into the detector with conservatively cropping facial regions like in [13]. All input images are resized to 299 × 299. For FF++, because the implementation of the face tracker used in [13] is not publicly available, we use the official manipulated region masks to localize facial regions. In addition, we calculate the labels of manipulation masks for FF++ and Celeb based on the differences between the fake faces and the real faces, like in [18]. Since it is hard to match manipulated faces to original faces in DFDC, we regard the facial regions in manipulated images as the corresponding manipulation masks.

4.2. Face forgery detection results

In previous works, the face forgery detectors are trained and evaluated under the same image quality degree. However, in practical application, the detectors have to deal with the images in unseen quality degrees. Thus, to properly evaluate models, we test trained detectors on seen and unseen image quality levels respectively.

4.2.1. Performances on seen image quality

We carry out experiments on FF++ to demonstrate the effectiveness of our methods. Because of the difference between our work and FF++ [13] in the face crop process, we re-implement three state-of-the-art detectors, i.e. Xception [13] and Face X-ray [6]., for fair comparison. We train models for all four manipulation methods at different image quality degrees separately. The accuracy of the binary detection task is shown in Table 1. On the raw images, all detectors obtain similar excellent performances, which benefits from the ample and significant details. On the HQ images, the performance of our model is slightly lower than the performance

Table 2
Comparison with previous state-of-the-art face forgery detectors in binary detection accuracy on LQ FF++. All models are trained for different manipulation methods separately. † represents our re-implementation.

Methods	DF	F2F	FS	NT
Steg. Features + SVM [20]	65.58	57.55	60.58	60.69
Cozzolino et al. [33]	68.26	59.38	62.08	62.42
Bayar and Stamm [34]	80.95	77.30	76.83	72.38
Rahmouni et al. [35]	73.25	62.33	67.08	62.59
MesoNet [15]	89.52	84.44	83.56	75.74
Xception [13]	94.28	91.56	93.70	82.11
Face X-ray† [6]	94.72	87.80	92.63	78.47
Xception† [13]	94.52	89.62	93.86	78.14
PRRNet (ours)	**95.63**	**90.15**	**94.93**	**80.01**

Table 3
Comparison of manipulation-specific forgery detectors in accuracy on the FF++. We train detectors on raw dataset and test them under different quality degrees respectively.

Methods	subset	raw	HQ	LQ
Xception [13]	DF	99.69	69.42	52.07
RPPNet	DF	**99.90**	**84.70**	**56.42**
Xception [13]	F2F	99.51	**93.79**	**52.59**
RPPNet	F2F	**99.90**	88.73	50.83
Xception [13]	FS	99.56	88.24	51.53
RPPNet	FS	**99.69**	**91.75**	**63.03**
Xception [13]	NT	98.04	65.00	**55.40**
RPPNet	NT	**98.11**	**77.72**	53.80

Table 4
Comparison between RPPNet and the state-of-the-art detector Xception in AUC on Celeb. We train and test detectors are on datasets in various qualities respectively.

Methods	train	raw	c20	c40
Xception [13]	raw	**99.81**	99.60	73.10
RPPNet	raw	99.80	**99.62**	**76.70**
Xception [13]	c20	99.67	99.61	78.85
RPPNet	c20	**99.79**	**99.71**	**79.37**
Xception [13]	c40	96.43	96.41	93.94
RPPNet	c40	**97.47**	**97.41**	**94.51**

Table 5
Comparison in AUC between RPPNet and the state-of-the-art detector Xception on DFDC. We report results for different augmentations separately.

Methods	original	low-fps	low-quality	low-res
Xception [13]	97.83	94.56	90.01	63.73
PRRNet	97.78	94.42	**91.75**	**82.39**

of re-implemented Xception. The one reason is the different backbones. The models with HRNet, both our model and the Face X-ray, have lower accuracy than Xception. However, our model performs better than Face X-ray with the same backbone, which shows the superiority of our method. The another reason is overfitting. The HQ images still retain enough details for face forgery detection. Thus, the simple network is qualified. And the complex model, such as ours, may cause slightly overfitting. However, our model demonstrates the remarkable superiority on the LQ images. When serious compression makes the low-quality images to be blurry, it is more difficult to distinguish the face forgeries only with visual features. While PRRNet achieves a new state-of-the-art with 86.13%. Compared to other detectors, PRRNet utilizes both the location information of the manipulated region and the spatial relation. The additional information could be evidence for determining the face forgeries when the visual features are influenced by the compression.

Then we evaluate PRRNet for each manipulation method independently on LQ images. The results are reported in Table 2. Noted, our Xception re-implemented by open-source code underperforms that in [13] for F2F and NT, although our Xception achieves similar or even better performances in most of the experiments. We argue the reason is that our preprocessing has some differences from that in [13]. The results show that PRRNet outperforms the re-implemented detectors by 0.91%, 0.53%, 1.07%, 1.54% for DF, F2F, FS, and NT respectively. This outperformance demonstrates the superiority of PRRNet. Besides, we observe that the outperformance of PRRNet evaluated for all four manipulation methods together is more than that when it is evaluated for each method independently. Due to the difference of manipulated methods, the features of face forgeries are discrepant. Hence, it is difficult for a fixed binary classifier to classify all face forgeries generated by different methods into one category accurately. In contrast, region-wise relations are similar in spliced facial manipulation. So PRRNet benefits from the region-wise relation and improves its capacity of detecting different kinds of face forgeries by the same kind of evidence.

4.2.2. Robustness to unseen image quality

The robustness of detectors to the change of the image quality is a meaningful property for face forgery detection. To verify the robustness of our PRRNet, we train PRRNet and test it under different qualities on three datasets respectively. In addition, we also employ Xception as a baseline.

Evaluate on FF++ In this experiment, both PRRNet and Xception are trained on the raw images, then they are tested on HQ images and LQ images respectively. Especially, we train all detectors for four different manipulation methods independently. Table 3 summarizes the results in accuracy of binary detection. We observe that the detectors reach excellent performances for raw images that are under the same quality degree as the train set. On HQ images, PRRNet outperforms Xception by 15.28%, 3.51%, and 12.72%, for DF, FS, and NT respectively. On LQ images, PRRNet outperforms Xception by 4.35% and 11.5%, for DF and FS respectively. It is demonstrated that our PRRNet is more robust than Xception to the change of image quality. However, in the rest cases, Xception reaches a little higher accuracy than PRRNet. We think there are two reasons. Firstly, the intra-image relation is also related to the visual information which is under the influence of the drop of quality. Secondly, the manipulated regions in F2F and NT images generally are very small, where it is difficult to capture their relation to original regions.

Evaluate on Celeb We also train PRRNet on Celeb under different image quality degrees respectively to evaluate its robustness. We report the result in terms of AUC (area under the Receiver Operating Characteristic curve) in Table 4. When trained on raw images, PRRNet and Xception perform similarly on raw images and c20 images, while PRRNet outperforms Xception on c40 images by 3.6%. In like manner, PRRNet also exceeds Xception by 0.52% on c40 images when trained on c20 images. Also, when trained on c40 images, PRRNet outperforms by 1.04% and 1% on raw and c20 respectively. These results demonstrate PRRNet is more robust than Xception.

Evaluate on DFDC The change of quality in DFDC is caused by three augmentations. Hence, based on the types of augmentations, the test set is divided into four subsets (one of them is from the original test set without any augmentation). In particular, the train set is not augmented. The results in terms of AUC are shown in Table 5. For original and low-fps images, PRRNet and Xception achieve similar results. While for low-quality images, PRRNet outperforms by 1.74%. The low-res images are under the lowest qual-

Table 6
Ablation study for different components of our proposed PRRNet on LQ NT subset of FF++ based on HRNet-w30.

Method	Relation Feature	Cosine Distance	Euclidean Distance	Inner Product	accuracy
HRNet					77.89
PRRNet	✓	✓			79.69
PRRNet	✓		✓		78.69
PRRNet	✓			✓	79.38
PRRNet		✓	✓	✓	78.30
PRRNet	✓	✓	✓	✓	**80.01**

ity, and PRRNet exceeds Xception by 18.66%. It indicates that PRRNet is robust to the change of image quality brought by various processes not only limited to compression.

Analyze the robustness To sum up, PRRNet is robust to the different image quality degrees. We argue that its robustness is benefit from exploiting the intra-image relation. It is constant that there are two parts from different sources in any spliced facial manipulated image. The image processing, such as compressing and downsampling, would change the image quality and break some important visual information. However, it hardly changes the relationship between the manipulated part and the original part. Therefore, the inconsistencies in manipulated images are a robust kind of evidence for face forgery detection.

4.3. Analysis of PRRNet

4.3.1. The ablation study of PRRNet

In this paragraph, we perform an ablation study to analyze the effect of different components in the proposed method. The two key factors of PRRNet are the pixel-wise relation and the region-wise relation. The former is captured by the PR module to generate the Relation Feature. While the latter is exploited by the RR module which fuse three metrics, i.e. Cosine Distance, Euclidean Distance, and Inner Product. Hence, we evaluate the performances of PRRNet under different setups on the LQ-NT subset of FF++. In particular, we employ HRNet which only contains a backbone network and a fully connected layer, as a baseline. The results are shown in Table 6. Except for the baseline, the models use the same pipeline as PRRNet: localizing the manipulated region and then detecting face forgery based on the intra-image relation. All these models outperform the baseline, which indicates that the location information and the relation are significant for face forgery detection. In addition, although the relation feature is expected to be mapped to the same feature space as the original local features, the bias between them is inevitable. Therefore, a complex metric is beneficial to measure the inconsistencies between features of different regions. So when relation feature is used, the PRRNet which fuses three metrics achieves a better result than other detectors that only employ a single metric. Moreover, without the relation features, the performance of PRRNet drops, which demonstrates the effect of the PR module.

4.3.2. The effect of the hyper-parameter n

In the PR module, we split the whole feature map into $n \times n$ patches. To evaluate the influence of the hyper-parameter n which determines the patch size, we train our PRRNet at $n = 3, 5, 7$ respectively. We use two convolutional layers without padding to map the split relation map R to relation feature whatever n is. The results are depicted in Table 7, where the model achieves the best performance at $n = 5$. On one hand, the bigger n is, the more information the relation map contains. On the other hand, the bigger n indicates that more parameters are required to extract features, which probably leads to overfitting. Therefore the value of n is a kind of trade-off. In fact, a model with a bigger n and more layers possibly achieves a better performance. However, it is not the key point in this work.

Table 7
The influence of the hyper-parameter n in the PR module.

n	3	5	7
accuracy	79.17	**80.01**	78.99

Table 8
The effect of the loss weight λ in the loss function.

λ	10	1	0.1	0.01
result	85.65	86.09	86.13	84.76

4.3.3. The effect of the loss weight λ

During training, we employ a hyper-parameter λ to balance L_{loc} and L_{cls} and set λ as 0.1 in the experiments. The results using different values of λ are shown in Table 8. We have three observations: (1) the performance when λ is set as 1 is comparable with accuracy being 86.09%; (2) when λ is set as 10, the result drops slightly with accuracy being 85.65%; (3) when λ is set as 0.01 the result is lower than others suggesting that the loss for measuring the inconsistencies is helpful.

4.3.4. Computational time

After demonstrating satisfactory performance on face forgery detection in the experiments, we evaluate the computational time of PRNet. The speed of PRRNet to forward propagate is 12.6 Frames Per Second (FPS) with batch size being 64 on four 1080Ti GPUs. Hence PRNet can detect manipulated images within one second second. To model the pixel-wise relation, the PR module calculates the similarity between every two pixels in the same feature map, which causes additional time consumption. Hence, in future work, we would try to optimize the time consumption of modeling the local relations.

5. How does PRRNet work?

5.1. Analysis for the pixel-wise relation

In this subsection, we visualize the similarities of all pixels to a given pixel respectively on the relation features generated by the PR module and the raw local feature before concatenating with relation features, in Fig. 5. Both similarity maps of two kinds of features are strongly related to the manipulated regions. The biggest difference between them is that the similarity maps of relation features have obvious boundaries, which indicates the relation features have more discriminant ability than the raw local features. Further, it is demonstrated that the PR module encodes the pixel-wise relation to relation feature space successfully.

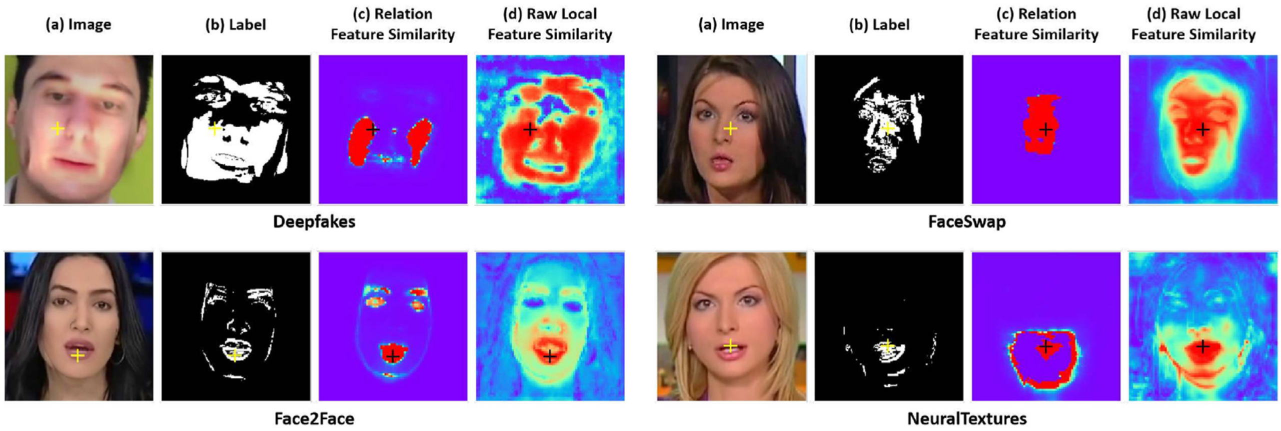
(a) Image
(b) Label
(c) Relation Feature Similarity
(d) Raw Local Feature Similarity
Deepfakes
FaceSwap
Face2Face
NeuralTextures

Fig. 5. Feature similarity visualization of all pixels to a given pixel marked by '+'. Red denotes high similarity.

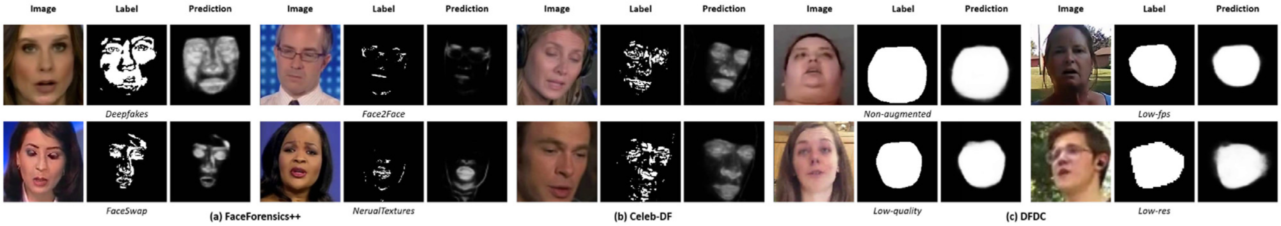

Fig. 6. Visual examples of face forgery localization.

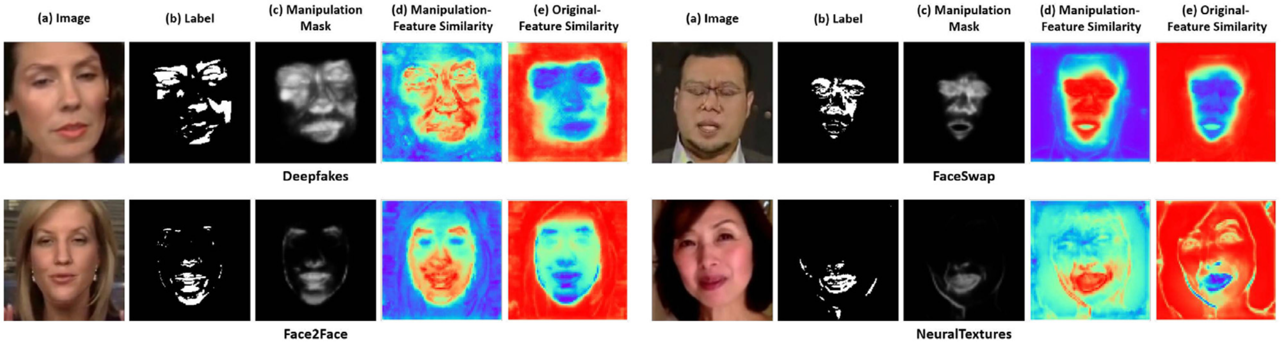

Fig. 7. Feature similarity visualization of features in all pixels to manipulation feature and original feature respectively. Column (a) shows the input images. (b) and (c) show the labels and predictions of manipulated regions respectively. (d) and (e) represent similarity maps of the manipulation feature and the original feature respectively.

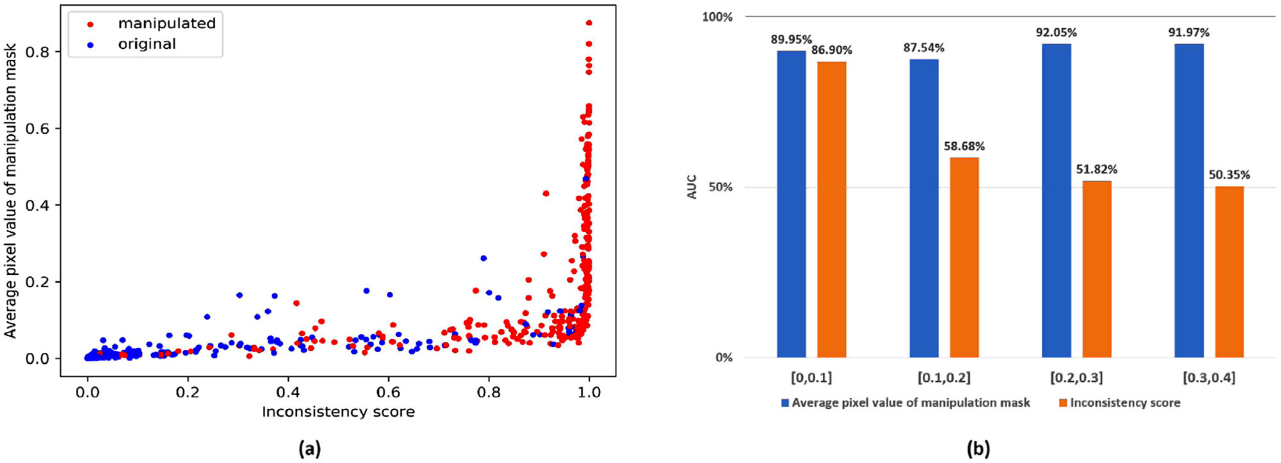

Fig. 8. (a) Visualization of the dependence of the RR module on the output of the PR module. (b) AUC on different subsets of LQ FF++. The AUC is computed respectively based on inconsistency scores and average pixel values of the manipulation masks.

5.2. Results in terms of forgery localization

Although determining whether the input images are manipulated is the primary goal in forgery detection, localizing the manipulated is also a meaningful task that could provide an interpretable kind of evidence for image forensics. Our PRNet can achieve the manipulated region localization with the manipulation mask M_{man} where each pixel represents the probability of the corresponding image pixels being manipulated. Some visual examples of manipulated images of various datasets are shown in Fig. 6. These examples demonstrate that our approach can locate the manipulated regions reliably. In addition, because the datasets use different ways to generate the labels of manipulated regions, the predictions show their characteristics correspondingly. For FF++ and Celeb, the predictions depend on the actual biases between the original faces and the manipulated faces. Therefore the predictions focus on the regions which have an obvious difference from original images. For DFDC, because we regard the whole face as the manipulated region, the predictions are seriously related to the facial regions.

5.3. Representations for regions

In this work, we obtain the region-wise features by aggregating the features of all positions with a spatial attention mechanism, as described in Section 3.2. Then to verify the effectiveness of this method, we visualize the similarity between the region-wise features and original feature maps in Fig. 7. It shows the features of different regions are indeed more similar to corresponding local features. Furthermore, this method which is based on the region-wise features is close to the human method, i.e. paying attention to local inconsistencies between different regions rather than only focusing on the global features of the images.

5.4. The significance of the RR module

The output of the PR module is the input of the RR module. Therefore, we attempt to illustrate the relation of the RR module on the PR module to demonstrate the significance of the RR module. Because the value of each pixel in the manipulation mask represents the probability of being manipulated, the average pixel value of manipulation mask should be zero for real images and be a large number for fake images. Hence, we evaluate the similarity between masks based on their average pixel value. For the sake of observation, we first normalize the average pixel values to [0,1]. Then, we visualize the relation of the inconsistency score computed by the RR module to the average pixel value of the manipulation mask, in Fig. 8 (a). We first observe that whatever the average pixel value is, there are inconsistency scores close to 1 for manipulated images. Moreover, when the average is close to 0, the RR module could correctly classify most of the test images. These observations demonstrate that the RR module learns more information about inconsistencies rather than completely depends on the outputs of the PR module.

We further analyze the relation with an experiment on LQ FF++. Because most of the images with average pixel values higher than 0.4 are manipulated, we divide the rest images into four subsets according to their average pixel values with 0.1 intervals. We calculate the AUC based on the average pixel value and the inconsistency score respectively for each subset, as shown in Fig. 8 (b). In the [0,0.1] subset, it is feasible to classify images only based on the average values. However, in other subsets, the average values of real and fake images are indistinguishable. In contrast, the inconsistency score shows its superior discriminability. Thus, modeling the region-wise relation to measure the inconsistencies is helpful for face forgery detection.

6. Conclusion

In this paper, a novel network PRRNet is proposed for face forgery detection. The PPRNet can achieve face forgery localization by exploiting the relation between the manipulated region and the original region in different levels. More specifically, the pixel-wise relation is captured to increases the discriminant ability of local features by encoding the feature similarity between every two pixels. Moreover, the region-wise inconsistencies is measured by multiple metrics to detect face forgery. Thus, our PRRNet exhibits excellent superiority on face forgery detection and shows good robustness on different image qualities. However, although our approach achieves promising performance on face forgery detection, our model is still restricted in some situations. For example, if an image is entirely synthetic, our method may hardly detect the manipulation by measuring inconsistencies. For future work, exploring the inter-frame inconsistencies in fake videos would be an interesting direction to further promote face forgery 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.

Acknowledgement

This work is supported by the National Key Research and Development Program of China (2017YFC0820600), the National Nature Science Foundation of China (62022076, U1936210), China Postdoctoral Science Foundation (2020M682035), the Youth Innovation Promotion Association Chinese Academy of Sciences (2017209), the Open Research Project of the State Key Laboratory of Media Convergence and Communication, Communication University of China (No. SKLMCC2020KF004).

References

[1] Deepfakes, (https://github.com/deepfakes/faceswap). Accessed: 2020-05-13.
[2] Faceswap, (https://github.com/MarekKowalski/FaceSwap). Accessed: 2020-05-13.
[3] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. Nießner, Face2face: Real-time face capture and reenactment of rgb videos, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2387–2395.
[4] J. Thies, M. Zollhöfer, M. Nießner, Deferred neural rendering: image synthesis using neural textures, ACM Transactions on Graphics (TOG) 38 (4) (2019) 1–12.
[5] H. Farid, Image forgery detection, IEEE Signal Process. Mag. 26 (2) (2009) 16–25.
[6] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, B. Guo, Face x-ray for more general face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5001–5010.
[7] A.C. Popescu, H. Farid, Exposing digital forgeries in color filter array interpolated images, IEEE Trans. Signal Process. 53 (10) (2005) 3948–3959.
[8] M.K. Johnson, H. Farid, Exposing digital forgeries by detecting inconsistencies in lighting, in: Proceedings of the 7th workshop on Multimedia and security, 2005, pp. 1–10.
[9] M. Chen, J. Fridrich, M. Goljan, J. Lukás, Determining image origin and integrity using sensor noise, IEEE Trans. Inf. Forensics Secur. 3 (1) (2008) 74–90.
[10] N. Liu, T. Zhou, Y. Ji, Z. Zhao, L. Wan, Synthesizing talking faces from text and audio: an autoencoder and sequence-to-sequence convolutional neural network, Pattern Recognit. 102 (2020) 107231.
[11] Z. Li, Y. Hu, R. He, Z. Sun, Learning disentangling and fusing networks for face completion under structured occlusions, Pattern Recognit. 99 (2020) 107073.
[12] Y. Fang, W. Deng, J. Du, J. Hu, Identity-aware cycleGAN for face photo-sketch synthesis and recognition, Pattern Recognit. 102 (2020) 107249.
[13] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Nießner, Faceforensics++: Learning to detect manipulated facial images, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 1–11.
[14] Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-DF: A large-scale challenging dataset for deepfake forensics (2020) 3207–3216.
[15] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, Mesonet: a compact facial video forgery detection network, in: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2018, pp. 1–7.
[16] R. Salloum, Y. Ren, C.-C.J. Kuo, Image splicing localization using a multi-task fully convolutional network (MFCN), J. Vis. Commun. Image Represent. 51 (2018) 201–209.

[17] H.H. Nguyen, F. Fang, J. Yamagishi, I. Echizen, Multi-task learning for detecting and segmenting manipulated facial images and videos, arXiv preprint arXiv:1906.06876 (2019).
[18] H. Dang, F. Liu, J. Stehouwer, X. Liu, A.K. Jain, On the detection of digital face manipulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5781–5790.
[19] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, C.C. Ferrer, The deepfake detection challenge (dfdc) preview dataset, arXiv preprint arXiv:1910.08854 (2019).
[20] J. Fridrich, J. Kodovsky, Rich models for steganalysis of digital images, IEEE Trans. Inf. Forensics Secur. 7 (3) (2012) 868–882.
[21] P. Ferrara, T. Bianchi, A. De Rosa, A. Piva, Image forgery localization via fine-grained analysis of CFA artifacts, IEEE Trans. Inf. Forensics Secur. 7 (5) (2012) 1566–1577.
[22] J. Lukas, J. Fridrich, M. Goljan, Digital camera identification from sensor pattern noise, IEEE Trans. Inf. Forensics Secur. 1 (2) (2006) 205–214.
[23] X. Yang, Y. Li, S. Lyu, Exposing deep fakes using inconsistent head poses, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 8261–8265.
[24] S. Bell, C. Lawrence Zitnick, K. Bala, R. Girshick, Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2874–2883.
[25] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, H. Lu, Dual attention network for scene segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3146–3154.
[26] X. Zeng, W. Ouyang, B. Yang, J. Yan, X. Wang, Gated bi-directional cnn for object detection, in: European conference on computer vision, Springer, 2016, pp. 354–369.
[27] Y. Liu, R. Wang, S. Shan, X. Chen, Structure inference net: Object detection using scene-level context and instance-level relationships, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6985–6994.
[28] Z.-M. Chen, X.-S. Wei, P. Wang, Y. Guo, Multi-label image recognition with graph convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5177–5186.
[29] H. Hu, J. Gu, Z. Zhang, J. Dai, Y. Wei, Relation networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3588–3597.
[30] F. Sung, Y. Yang, L. Zhang, T. Xiang, P.H.S. Torr, T.M. Hospedales, Learning to compare: Relation network for few-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1199–1208.
[31] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, J. Wang, High-resolution representations for labeling pixels and regions, arXiv preprint arXiv:1904.04514 (2019).
[32] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[33] D. Cozzolino, G. Poggi, L. Verdoliva, 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, 2017, pp. 159–164.
[34] B. Bayar, M.C. Stamm, 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, 2016, pp. 5–10.
[35] N. Rahmouni, V. Nozick, J. Yamagishi, I. Echizen, Distinguishing computer graphics from natural images using convolution neural networks, in: 2017 IEEE Workshop on Information Forensics and Security (WIFS), IEEE, 2017, pp. 1–6.

Zhihua Shang received the B.E. degree in electronic information engineering at Northwestern Polytechnical University, Xi'an, China, in 2015. He is currently pursuing the Ph.D. degree at School of Information Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include face forgery detection and localization.

Hongtao Xie received the Ph.D. degree in computer application technology from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2012. He is currently a Professor with the School of Information Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include multimedia content analysis and retrieval, deep learning, and computer vision.

Zheng-Jun Zha (M'08) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2004 and 2009, respectively. He is currently a Full Professor with the School of Information Science and Technology, University of Science and Technology of China, the Vice Director of National Engineering Laboratory for Brain-Inspired Intelligence Technology and Application. He was a Researcher with the Hefei Institutes of Physical Science, Chinese Academy of Sciences, from 2013 to 2015, a Senior Research Fellow with the School of Computing, National University of Singapore (NUS), from 2011 to 2013, and a Research Fellow there from 2009 to 2010. His research interests include multimedia analysis, retrieval and applications, as well as computer vision etc. He has authored or coauthored more than 100 papers in these areas with a series of publications on top journals and conferences. He was the recipient of multiple paper awards from prestigious multimedia conferences, including the Best Paper Award and Best Student Paper Award in ACM Multimedia, etc. He serves as an Associated Editor of IEEE Trans. on Circuits and Systems for Video Technology.

Lingyun Yu is currently an Associate Professor with Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, and a Postdoc with University of Science and Technology of China. She received her B.S. degree from China University of Mining and Technology, in 2015, and her Ph.D. degree from University of Science and Technology of China, in 2020. Her research interests cover talking face generation, face forgery detection, multi-modal learning, articulatory movements-driven 3D talking head and video synthesis.

Yan Li received the master degree in computer science from Institute of Computing Technology, Chinese Academy of Sciences in 2012. He is currently the head of Multimedia Understanding (MMU) Department of Kuaishou Technology, and a member of the Multimedia Technology Standing Committee in China Computer Federation (CCF). His current research interests are in the fields of computer vision and multimedia content analysis.

Yongdong Zhang (M'08-SM'13) received the Ph.D. degree in electronic engineering from Tianjin University, Tianjin, China, in 2002.He is currently a Professor with the School of Information Science and Technology, University of Science and Technology of China. His current research interests are in the fields of multimedia content analysis and understanding, multimedia content security, video encoding and streaming media technology. He has authored over 100 refereed journal and conference papers. He was a recipient of the Best Paper Awards in PCM 2013, ICIMCS 2013, and ICME 2010, the Best Paper Candidate in ICME 2011.He serves as an Associate Editor of IEEE Trans. on Multimedia and an Editorial Board Member of Multimedia Systems Journal.