IR-capsule: two-stream network for face forgery detection.

LIN, K., HAN, W., LI, S., GU, Z., ZHAO, H., REN, J., ZHU, L. and LV, J.

2022

IR-Capsule: Two-Stream Network for Face Forgery Detection

Kaihan Lin · Weihong Han * · Shudong Li * · Zhaoquan Gu · Huimin Zhao · Jinchang Ren · Li Zhu · Jujian Lv

*** Corresponding author. Email: hanweihong@gzhu.edu.cn; lishudong@gzhu.edu.cn**

Received: date/ Accepted

Structural abstract

Background

With the emergence of deep learning, generating forged images or videos has become much easier in recent years. Face forgery detection, as a way to detect forgery, is an important topic in digital media forensics. Despite previous works having made remarkable progress, the spatial relationships of each part of the face that has significant forgery clues are seldom explored.

Methods

To overcome this shortcoming, a two-stream face forgery detection network that fuses Inception ResNet stream and capsule network stream (IR-Capsule) is proposed in this paper, which can learn both conventional facial features and hierarchical pose relationships and angle features between different parts of the face. Furthermore, part of the Inception ResNet V1 model pre-trained on the VGGFACE2 dataset is utilized as an initial feature extractor to reduce overfitting and training time, and a modified capsule loss is proposed for the IR-Capsule network.

Results

Experimental results on the challenging FaceForensics++ benchmark show that the proposed IR-Capsule improves accuracy by more than 3% compared with several recently published methods.

Conclusions

The proposed method provides a new solution for face forgery detection, which has outperformed a few state-of-the-art models.

Keywords

Two-stream network · Face forgery detection · IR-Capsule · Capsule network · Inception ResNet.

Kaihan Lin · Weihong Han · Shudong Li · Zhaoquan Gu

Cyberspace Institute of Advanced Technology, Guangzhou University, Guangzhou, China

* Weihong Han and Shudong Li are corresponding authors. Email: hanweihong@gzhu.edu.cn; lishudong@gzhu.edu.cn

Weihong Han

Peng Cheng Laboratory, Shenzhen, China

Huimin Zhao · Jinchang Ren · Jujian Lv

School of Computer Science, Guangdong Polytechnic Normal University, Guangzhou, China

Jinchang Ren

National Subsea Centre, Robert Gordon University, Aberdeen, U.K

Li Zhu

Industrial Training Center, Guangdong Polytechnic Normal University, Guangzhou, China

1 Introduction

Human face is one of the most representative and recognizable features among biometric features. Therefore, facial features have been successfully applied in various application fields [1-3]. However, the security threat posed by face forgery is increasing with the rapid development of face-related online authentication applications [4-5]. Face forgery refers to the technology of synthesizing faked facial images or videos using methods such as deep learning. It can superimpose the image of the target person's face to the corresponding position of the original person's face in the

video or change the expression and pose of the target person in a way as if the target person is saying some specific words or even sentence, or does some faked actions [6]. This technology can also be used to create virtual characters in movie production and resurrect historical figures or dead relatives and friends in videos. Overall, its disadvantages still outweigh its advantages. Face forgery often contains faked characters, events and voice information, such as fake news, Internet fraud, rumors, etc. [7-9], which can negatively impact our lives. Face forgery detection has become an important subject to be studied urgently because the abuse of face forgery technology has brought great threats to personal privacy and social security.

In recent years, with the development of deep learning technology, a large number of face forgery detection approaches based on convolutional neural network (CNN) have been proposed. However, traditional CNN has some limitations in the field of image processing, such as not considering the relative position, angle and other information between components and poor interpretability. As for face forgery, attackers tend to modify facial features, leaving behind hard-to-detect clues such as the relative positions and angles of various parts of the face [10]. Traditional CNN-based detection methods are difficult to detect these important forged clues due to the above limitations, which will affect the detection performance. A remarkable solution to the limitation problem of traditional CNN is the capsule network based on dynamic routing proposed by Hinton et al. [11]. The capsule network is a novel network architecture, which improves the input of neurons from scalar to vector, and adjusts network parameters through dynamic routing to integrate features. Vector representation greatly enriched the expression ability of features, so that the capsule network can obtain the relative position, angle and other information between the components of the image. With these advantages, capsule network has been successfully applied to some fields of computer vision [12-14].

To effectively detect the fake face generated by the state-of-the-art forgery algorithms and overcome the shortcomings of traditional CNN, a two-stream network called IR-Capsule for face forgery detection is proposed in this paper. One stream of our network is the Inception ResNet for conventional facial feature extraction, and the other is the capsule network for the extraction of location-related features such as relative positions and angles between facial parts. Our approach combines these two types of features, and the information gained from different networks is shared among them, thereby the network's overall performance has been much improved.

The main contributions of this paper are highlighted as follows: 1) A two-stream network for face forgery detection is proposed, which can learn the conventional facial features and the hierarchical pose relationships and angle features between different parts of the face; 2) One part of the Inception ResNet V1 model pre-trained on VGGFACE2 face recognition dataset is used as an initial feature extractor, which can mitigate overfitting and reduce the training time of the model; 3) A modified capsule loss of the capsule network is proposed to make it suitable for our two-stream network, and the idea of ensemble learning is also applied for improving the final output.

The remainder of this paper is organized as follows. Section 2 briefly reviews the related work of face forgery detection. The IR-Capsule framework for face forgery detection and its key aspects are described in detail in Section 3. Section 4 presents the experimental results and discussion of the proposed IR-Capsule. Finally, Section 5 summarizes this work and discusses the direction of the future work.

2 Related work

2.1 Face forgery generation

Interest in face forgery has increased rapidly recently. As early as 1997, Bregler et al. [15] proposed a video rewrite method by generating mouth movements to automatically create new videos containing fake faces. Since then, face forgery technology has developed rapidly, especially in the past decade. Alexander et al. [16] scanned an actress's 33 facial expressions through sophisticated equipment to synthesize a digital version of her. Dale et al. [17] proposed a method for replacing facial expression in videos, which took into account the differences in identity, visual appearance, speech and time of source and target videos. Garrido et al. [18] proposed a system for modifying the lip motion of an actor in a video to match the video to the target audio track. Thies et al. [19] proposed the first real-time expression transfer method for facial reenactment, and on this basis, a self-reenactment method for eye tracking and representation in virtual reality was proposed [20] and extended to the full-body [21]. In [22], a method to generate and modify the facial expressions of target actors was

learned while maintaining their styles through a recurrent generative adversarial network (GAN) [23]. Nirkin et al. [24] presented a FSGAN algorithm based on GAN for face swapping and reenactment, which did not require the training of new faces. Tripathy et al. [25] proposed a generic face animator that could control the pose and expressions of a given face image.

The above face forgery generation methods are more inclined to academic research, which require certain professional knowledge and equipment, so they are not widely used in real scenes. For practical applications, the current state-of-the-art face forgery generation methods include computer graphics-based FaceSwap [26] and Face2Face [27], and deep learning based DeepFakes [28] and NeuralTextures [29]. FaceSwap achieves facial identity forgery by smoothly transferring the facial region, a light-weight approach that can be implemented efficiently on the CPU. Face2Face is a graphics-based facial reenactment system that can maintain the identity of the target, but transfer facial expressions from the source. DeepFakes is based on auto-encoder structure training to reconstruct the source and target faces, and NeuralTextures is based on GAN to modify the mouth region. Several examples of fake faces generated by these four face forgery generation methods are shown in Figure 1, which shows that it is not easy to tell the real from the fake. Increasing advanced face forgery technology makes it more difficult for people to distinguish between real and fake, bringing severe trust issues and security risks to our society. Therefore, it is crucial to study an effective face forgery detection method.

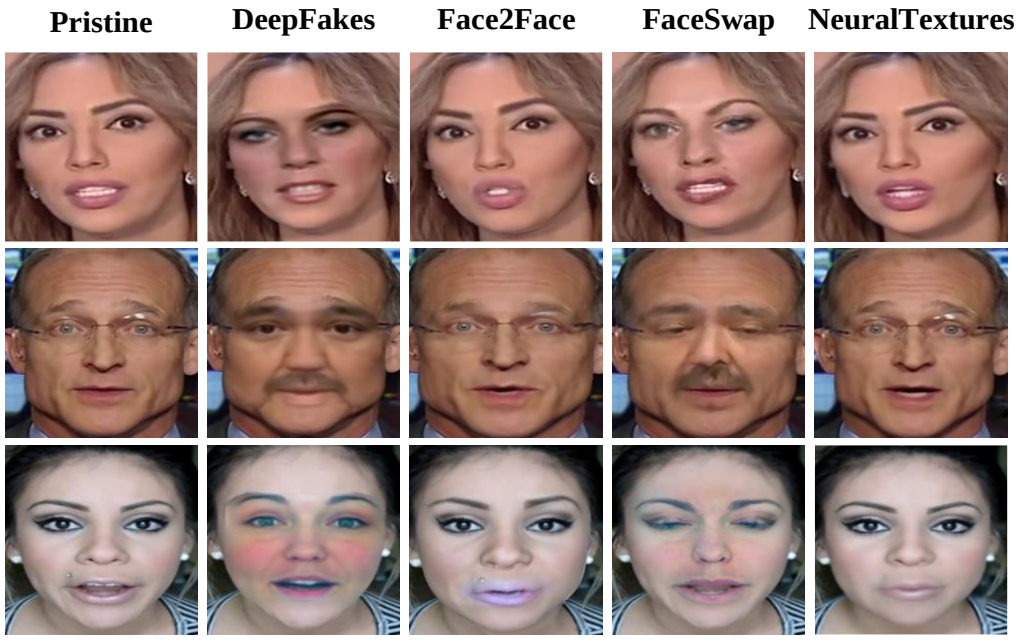

Fig. 1 Examples of fake faces from the FaceForensics++ dataset. The first column is the pristine image, and the second to fifth columns are the forged images generated by DeepFakes, Face2Face, FaceSwap, and NeuralTextures.

2.2 Face forgery detection

Face forgery detection, as an effective approach to defend against forgery attacks, has received increasing attention. As suggested in [10], we classify face forgery detection approaches into conventional approaches and deep learning based approaches.

Conventional approaches: Early attempts can detect faces forgery according to the image features that researchers target before the emergence of deep learning. Fridrich and Kodovsky [30] proposed steganalysis-based detectors for forgery detection, which assembled the rich model of noise components into the union of many different sub-models and formed the joint distribution of adjacent samples of quantified image noise residue. Subsequently, an improved version of the approach was implemented in CNN [31]. Lyu et al. [32] proposed a method to detect forgery by revealing the inconsistency between global noise and local noise, which was based on the fact that images from different sources may have different noise characteristics introduced by sensors or post-processing steps. Another practical approach was to utilize color filter array (CFA) model analysis for forgery detection distinguishing the tampered regions and the authentic regions [33-37]. Compression artifacts were also the key solution for forgery detection, which could be classified according to the clues they rely on. Fan et al. [38] were pioneers in applying block artifact grid (BAG) to digital media forensics, which utilized

JPEG block processing to form a grid pattern that is easy to detect. Since then, BAG-based approaches were proposed in [39-42]. In general, conventional detection approaches have the advantages of speed and extendibility, but the detection accuracy is limited compared to deep learning based methods.

Deep learning based approaches: Recently, witnessing the impressive achievements of deep learning in the field of computer vision, researchers have gradually applied deep learning to forgery detection. Zhou [43] et al. proposed a two-stream network [44] for face forgery detection, in which one network was leveraged to detect tamper artifacts, and the other network was leveraged to capture local noise residuals and camera characteristics. In [45-46], a recurrent neural network was introduced into the field of forgery detection to classify the authenticity of each frame of video. Nguyen et al. proposed some effective forgery detection methods [47-48] and pioneered the application of capsule networks to digital media forensics [49-50]. A different view from other methods was proposed in [51], which used the optical flow fields to exploit possible inter-frame differences. Rossler et al. constructed a well-known face forgery dataset FaceForensics++ and utilized the Xception network for face forgery detection in [52]. Transfer learning was also utilized for forgery detection and took into account the temporal information of the video in [53]. Li et al. [54] proposed a method called face X-ray that focuses on the face boundaries to observe the blending operation traces that occurred during forgery. The method based on deep learning had high effectiveness and stability, so it became the mainstream of face forgery detection. However, the existing deep learning based approaches mainly focused on the conventional facial features and ignored the hierarchical pose relationships of each part of the face, which may limit the detection accuracy. In this paper, we proposed a new solution called IR-Capsule, which comprehensively considered the traditional features and the hierarchical structure relationship of each part of the face. The experimental results have validated that the IR-Capsule network achieves promising performance compared to state-of-the-art methods.

3 IR-Capsule Network

As illustrated in Figure 2, our IR-Capsule is composed of two stream networks, one of which is the Inception ResNet network to extract conventional facial features, and the other is the capsule network to capture the hierarchical pose relationships between facial components. The first step of the IR-Capsule is to carry out face detection on the input image. In this phase, we adopt a multi-task cascaded convolutional network (MTCNN) [55], since it can quickly and efficiently complete face detection and face alignment. Then, unlike other methods that deal with the complete image directly, we crop the image to only contain the facial area. This process reduces the interference of other information in the image, which can speed up training and improve performance. In addition, to ensure that the cropped image is large enough and contain sufficient information, we adopted the image size of 300×300 commonly used in practice. We did not use a larger image size, since it will bring greater computational cost. The cropped image is then fed into the initial layers of the Inception ResNet V1 network, which has been pre-trained on the VGGFACE2 [56] dataset. The VGGFACE2 dataset is one of the largest face recognition datasets available, so pre-training Inception ResNet V1 network in this dataset can provide more efficient face feature extraction, reduce training time and avoid overfitting [57]. In the last part, a two-stream network is utilized to detect the traces of face forgery, and the final output is obtained through the fusion of the two streams. The detailed architecture of Inception ResNet and capsule network streams are discussed in Section 3.1 and 3.2 respectively.

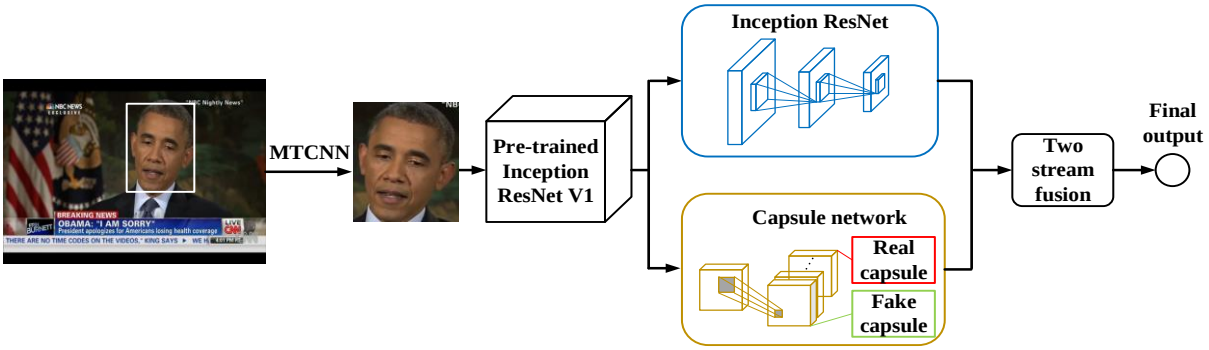

Fig. 2 Network structure of IR-Capsule.

3.1 Inception ResNet Stream

IR-Capsule is a two-stream network, which includes an Inception ResNet stream and a capsule network stream. In the Inception ResNet stream, the complete Inception ResNet V1 is used to extract the deep conventional features after the initial feature extractor. For Inception ResNet series networks, we chose the Inception ResNet V1 network because its improved version has higher model complexity and requires more computational resources. While face forgery detection is a binary classification task, Inception ResNet V1 can provide sufficient performance with low computational power consumption. The details of the Inception ResNet stream are shown in Figure 3. It can be seen that Inception ResNet V1 is composed of multiple blocks including Inception-resnet, which has the benefits of Inception and ResNet. Specifically, part of the pre-training network consists of several initial layers of the Inception ResNet V1 network, which are not too deep and do not have a significant impact on subsequent training.

3.2 Capsule Network Stream

The "capsule" in the capsule network is a set of vectors or matrices, where different output vectors represent different attributes of specific objects appearing in the image. The capsule network uses vector output instead of scalar output, recognizing the object category while retaining its hierarchical posture relationship and angle information. In addition, the dynamic routing algorithm is utilized to update the weights from low-level to high-level instead of pooling, thereby avoiding information loss. There may be some abnormality in the hierarchical posture relationship and angle between various parts of the face for face forgery. Therefore, the capsule network can effectively detect traces of forgery compared with the conventional neural networks.

The capsule network stream is designed after the pre-trained network, extracting deeper features such as hierarchical pose features and angle features. As stated in Figure 4, the capsule network stream mainly consists of ten primary capsules and two output capsules (details of parameter settings can be found in Section 4.2). The ten primary capsules have the same design, and statistical pooling is introduced in the middle part. For the next part of the network, we use the real capsule and the false capsule as the output capsule, which is suitable for the binary classification task of face forgery detection. The dynamic routing algorithm is used to update the weight between the primary capsule and the output capsule. After the normalization process, the capsule network stream output the prediction result.

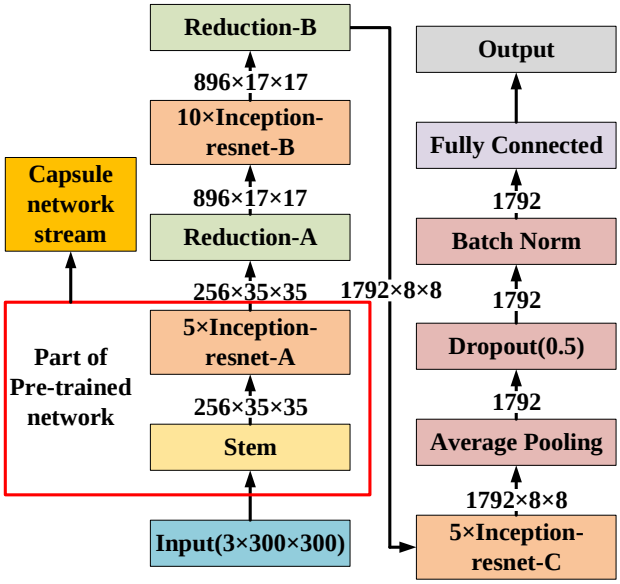

Fig. 3 Details of the Inception ResNet stream.

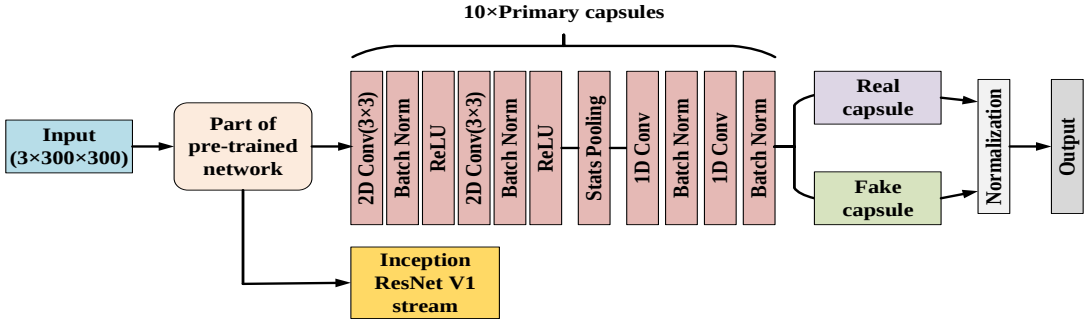

Fig. 4 Details of the capsule network stream.

The dynamic routing algorithm is used to calculate the weight between the primary capsule and the output capsule, routing these primary capsules to the output capsule in real time. Therefore, the output capsule can be used to express the consistency of low-level capsule features. Let u_i and v_j represent the output vector of the primary capsule and the output capsule, respectively. The input s_j of the capsule j can be calculated by the coupling coefficient $c_{i,j}$ determined in the iterative dynamic routing process as follows:

$$s_j = \sum_i c_{i,j} \hat{u}_{j|i}, \quad \hat{u}_{j|i} = W_{ij} u_i, \tag{1}$$

where W_{ij} is the weight matrix and $\hat{u}_{j|i}$ is the weighted summation result of all prediction vectors. The dynamic routing algorithm adjusts the coupling coefficient $c_{i,j}$ through several iterations to determine the appropriate weight, and the coupling coefficient $c_{i,j}$ can be calculated as follows:

$$c_{i,j} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})}, \tag{2}$$

where the initial logits b_{ij} are the log prior probabilities that the primary capsule i should be coupled to the output capsule j. Then s_j is compressed by a squash function to obtain the output vector v_j of the output capsule by

$$v_j = squash(s_j) = \frac{\|s_j\|^2}{1+\|s_j\|^2} \frac{s_j}{\|s_j\|}. \tag{3}$$

The overall process of the dynamic routing algorithm is described in Table 1. To reduce overfitting, we also add the random noise to W_{ij} and then use a dropout operation. We set up three iterations ($r = 3$) through repeated experiments, for a trade-off between the accuracy and speed.

For the output of the capsule network stream, different from the softmax function used for the output vector of the output capsule in [11, 49], we have done a simple processing for the output vector of the output capsule. Let x be the output vector of the real capsule, the square root of its square value is the final output. Let denoted as $v_{real} = \sqrt{x^2}$. Therefore, the capsule loss function can make the magnitude of output no less than 0.9 when the input image is real, and no greater than 0.1 when the input image is fake. The details of the capsule loss function are described in Section 3.3.

Table 1. Dynamic routing algorithm

Procedure 1 Dynamic routing algorithm.

procedure ROUTING(u_i , r)

for all input capsule i and all output capsules j : $b_{i,j} \leftarrow 0$

for r iterations do

for all input capsules i : $c_i \leftarrow softmax(b_i)$

for all input capsules j : $s_j \leftarrow \sum_i c_{i,j} \hat{u}_{j|i}$

for all input capsules j : $v_j \leftarrow squash(s_j)$

for all input capsules i and output capsules j : $b_{i,j} \leftarrow b_{i,j} + \hat{u}_{j|i} v_j$

return v_j

3.3 Two Stream Fusion

The IR-Capsule network combines Inception ResNet stream and capsule network stream to detect face forgery. For Inception ResNet stream, the sigmoid function is utilized to normalize its output. Subsequently, the binary cross entropy function is utilized as the loss function, which is defined as follows:

$$L_{IR} = -\frac{1}{m}\sum_{i=1}^{m} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i), \tag{4}$$

where $\hat{y}_i$ is the final output after the sigmoid function, and y_i is the ground-truth label of i-th input. In the training process, we define 1 as a fake face and 0 as a real face.

For the capsule network stream, unlike the cross entropy loss function used in [49], we simply improved the traditional capsule loss [11]. We only utilized the margin loss and discarded the reconstruction loss, which can be more suitable for our network. In addition, this loss function can make one output capsule output a vector of magnitude 0.9 or greater, and the other output is equal to or less than 0.1 during the training process. The loss function of capsule network stream is defined as follows:

$$L_c = \frac{1}{m}\sum_{i=1}^{m} T_i \max(0, 0.9 - v_i) + \lambda(1-T_i)\max(0, v_i - 0.1), \tag{5}$$

where T_i is the ground-truth label of i-th inputs, and v_i is the predicted probability of the capsule stream. For the weight value λ, we set it as 0.5 following [11]. This loss would encourage the IR-Capsule network to match the index of the "correct" capsule with the true class label of given

training data. Finally, the total loss function is the sum of two streams, which is defined as follows:

$$L = L_{IR} + L_C . \quad (6)$$

The final output is a fusion of the two streams. In the test process, the final output is the probability that the input face image is a forged image. The input image is fake if the probability of the final output P_F is greater than 0.5, otherwise it is a real image. We refer to the idea of ensemble learning and set balance factor λ=0.5 following [58]. The final output is defined as follows:

$$P_F = \lambda P_C + (1-\lambda) P_{IR} , \quad (7)$$

where P_C is the output probability of capsule network stream and P_{IR} is the output probability of Inception ResNet stream.

4 Experiments

In this section, we first introduce the overall experiment setup, including dataset and implementation details. Then, the main parameters of IR-Capsule are discussed. Finally, we demonstrate the effectiveness of our method compared with state-of-the-art methods.

4.1 Experimental setting

To effectively evaluate the IR-Capsule network, the well-known FaceForensics++ dataset is leveraged in the training and testing process. It is a large and challenging dataset, including the fake faces generated by the four state-of-the-art forgery generation methods of FaceSwap, DeepFakes, Face2Face, and NeuralTextures. The FaceForensics++ dataset collects 1000 original videos from YouTube, and each video is generated by the above four forgery generation methods to generate 4000 fake videos, a total of 5000 videos. In addition, each video has three different compression rates (no compression: raw, medium compression: c23, high compression: c40), and the number of frames is between 300 and 700. In the process of training and testing, we followed [52] to divide each group of 1000 videos into 720 for training, 140 for validation and 140 for testing. Since face forgery detection can be defined as a binary classification task, we use binary labels in the training process. Moreover, we took the second 100 frames of the video because we found that the first 100 frames of the input video might be unstable.

All experiments were performed on the same device during the implementation process, which was Intel(R) Xeon(R) Bronze 3206R CPU of 32GB flash memory and NVIDIA GeForce GTX 2080 Ti GPU. The PyTorch 1.7.0 library was utilized for training and testing on the Ubuntu 18.04 OS. The IR-Capsule network was trained for 25 epochs through the Adam optimizer. We also empirically set the learning rate to 0.0001, batch size to 64, and the β_1 and β_2 in the Adam optimizer to 0.9 and 0.999 respectively

4.2 Parameter analysis

There are several tuning parameters in our proposed IR-Capsule network. We discussed the effect of two main parameters on the performance, which is the number of primary capsules i and the number of iterations r. The detection accuracy of different typical parameter settings in the FaceForensics++ testing set is reported in Table 2. The first six rows show the results for $r = 3$ and the last three rows show the results for $i = 3$. As seen in the first six rows, the setting $i = 10$ has the best performance, while $i = 3$ has the worst. The detection accuracy increased with the number of capsules, but did not improve when $i > 10$. Experiments demonstrated that a reasonably large number of primary capsules may improve network performance. For the last three rows, we set the number of primary capsules $i = 10$ to compare the accuracy under different iterations. We can see that the results improve slightly as the number of iterations increases. However, when the r is increased to 4 and 5, no improvement can be found when compared to $r = 3$. That is because as the number of iterations r increases, the model will overfit the training set, thus affecting its performance on the testing set. According to the experimental results of parameter analysis, we have selected the optimal parameter $r = 3$ and $i = 10$.

Table 2. Detection accuracy of different typical parameter settings.

Iterations r	Primary capsules i	Accuracy (%)
3	3	80.24
3	5	87.19
3	8	89.67
3	10	**91.03**
3	16	90.81
3	32	88.36
2	10	89.21
4	10	89.77
5	10	90.63

4.3 Comparison with other approaches

We demonstrate our IR-Capsule network on the well-known FaceForensics++ benchmark and compare the results with

several state-of-the-art approaches. The FaceForensics++ benchmark is a challenging face forgery detection benchmark, which contains 1000 unseen images. In these 1,000 unseen images, each picture is randomly extracted from the forgery generation method or the pristine video, including a random compression level. The ground truth labels are hidden and the submitted approaches are automatically evaluated online for their classification accuracy. Since the FaceForensics++ benchmark contains these 1000 videos that are different from training and validation data, the generalization and effectiveness of the submitted approaches can be evaluated.

We first compared 10 published state-of-the-art approaches in the FaceForensics++ benchmark, including Steganalysis Features [30], Recasting [31], Rahmouni [59], Bayar and Stamm [60], p-DARTS [61], XceptionNet Full Image [52], GAEL-Net [62], MesoNet [63], XceptionNet [52], and Inception ResNet V1 [53]. The experimental results are shown in Table 3. It can be seen that our method has achieved promising performance, and the overall performance surpasses all the comparison approaches. In detail, our method is 2.2% higher than Inception ResNet V1 on FaceForensics++ benchmark, 12.1% higher than XceptionNet, and 17.1% higher than MesoNet. In addition, the performance of the IR-Capsule network in pristine image classification is better than all comparison methods.

Table 3. Experimental results on FaceForensics++ benchmarks. We report the precision results of four forgery methods and pristine images, as well as the overall accuracy.

Detection Methods	Accuracy					
	DeepFakes [28]	Face2Face [27]	FaceSwap [26]	NeuralTextures [29]	Pristine	Total
Steganalysis Features [30]	0.736	0.737	0.689	0.633	0.340	0.518
Recasting [31]	0.855	0.679	0.738	0.780	0.344	0.552
Rahmouni [59]	0.855	0.642	0.563	0.607	0.500	0.581
Bayar and Stamm [60]	0.845	0.737	0.825	0.707	0.462	0.616
p-DARTS [61]	0.791	0.730	0.816	0.720	0.478	0.618
XceptionNet Full Image [52]	0.745	0.759	0.709	0.733	0.510	0.624
GAEL-Net [62]	0.718	0.686	0.631	0.707	0.562	0.625
MesoNet [63]	0.873	0.562	0.612	0.407	0.726	0.660
XceptionNet [52]	0.964	**0.869**	0.903	0.807	0.524	0.710
Inception ResNet V1 [53]	0.936	0.839	0.903	**0.820**	0.750	0.809
Proposed IR-Capsule	**0.973**	0.818	**0.942**	0.793	**0.792**	**0.831**

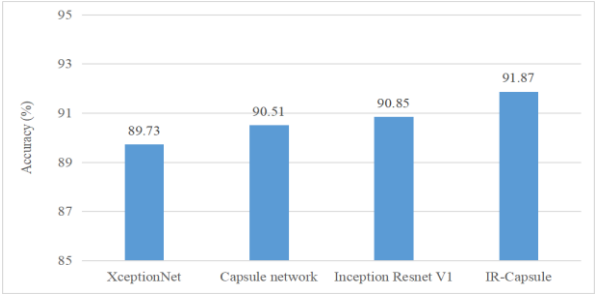

Fig. 5 Experimental results on FaceForensics++ validation set.

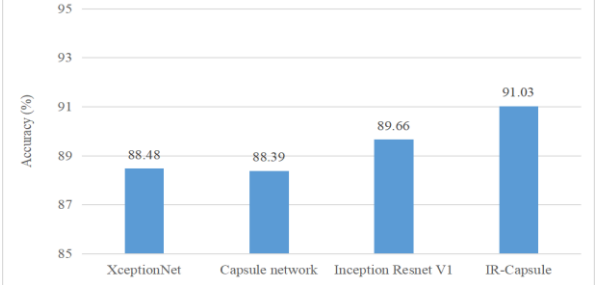

Fig. 6 Experimental results on FaceForensics++ testing set.

To further demonstrate the performance of IR-Capsule network, we compared it with XceptionNet, Inception ResNet V1 and Capsule network [49] on the entire validation and testing sets of FaceForensics++. As shown in Figure 5 and Figure 6, the IR-Capsule network has the highest accuracy in both the validation and testing sets. We also performed a deeper analysis of the detection performance of different face forgery methods, including FaceSwap, DeepFakes, Face2Face, NeuralTextures. The model is trained on all forgery methods data and evaluated on specific forgery method. The detection accuracy of different forgery methods in the validation and testing sets are shown in Figure 7 and Figure 8. It can be seen that our method outperforms state-of-the-art methods in the detection of different types of face forgery.

4.4 Comparison of computational time

To evaluate the efficiency of the proposed IR-Capsule, we have compared in Table 4 the computational times of various methods with 100 randomly selected forged images. Meanwhile, as an indicator of computational complexity, the number of parameters of each method is also reported in Table 4. As seen in Table 4, the number of parameters and computational time of our method is comparable to Inception ResNet V1 and superior to the XceptionNet approach. Although the number of parameters of the capsule networks is minimal, the total computation time is only 0.043s faster than our method. This is because the pre-processing and face detection process before entering the capsule network takes most of the computational time. In general, the computational time of the proposed IR-Capsule approach is at the same level as other detection approaches, but our approach has a higher detection accuracy. The experimental results of detection performance and computational time fully demonstrate the efficiency and efficacy of the proposed IR-Capsule framework.

Table 4. The number of parameters and computational time of different detection methods.

Detection Method	Number of parameters	Computational time (s)
XceptionNet (299×299)	27,910,840	0.671
Inception ResNet V1(300×300)	23,483,137	0.493
Capsule network (300×300)	**4,176,574**	**0.455**
IR-Capsule (300×300)	24,954,207	0.498

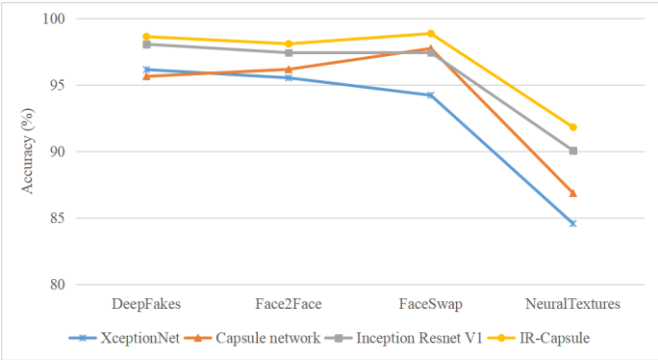

Fig. 7 Performance comparison of different forgery methods on FaceForensics++ validation set.

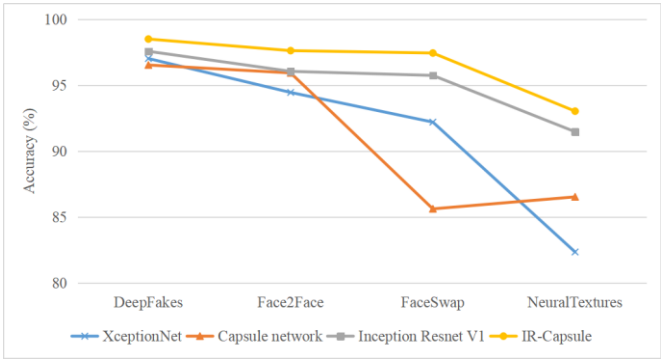

Fig. 8 Performance comparison of different forgery methods on FaceForensics++ testing set.

5 Conclusion

In this paper, a two-stream network that fuses Inception ResNet stream and capsule network stream (IR-Capsule) is proposed for face forgery detection, which aims to overcome the shortcomings of traditional CNN-based detectors, especially the lack of attention to the spatial relationships of each part of the face. In the proposed approach, the Inception ResNet stream is utilized for conventional feature extraction, and the capsule network stream is utilized to extract relative position and angle features between face parts. In addition, an improved capsule loss function is leveraged to replace the original capsule loss, and a part of the Inception ResNet V1 model pre-trained on the VGGFACE2 face recognition dataset is leveraged as the feature extractor before the IR-Capsule network. The experimental results show that our approach outperforms the state-of-the-art approaches as its ability to learn both conventional facial features and forged artifacts of relative position and angle. Future work will consider more effective features to further improve the accuracy of face forgery detection.

Acknowledgments

This work was partly supported by National Natural Science Foundation of China (No. 61972106, U1803263, 61902082), National Key research and Development Plan (Grant No. 2019QY1406), Key-Area Research and Development Program of Guangdong Province (Grant No. 2019B010136003), Dongguan Innovative Research Team Program (No. 2018607201008), Guangdong Higher Education Innovation Group (No. 2020KCXTD007), Guangzhou Higher Education Innovation Group (No. 202032854), Key Laboratory of the Education Department of Guangdong

Province (No. 2019KSYS009), Scientific and Technological Planning Projects of Guangdong Province(No. 2021A0505030074).

Compliance with Ethical Standards

Conflict of Interests

The authors declare that they have no conflict of interest.

References

[1] Deng J, Guo J, Xue N, Zafeiriou S. Arcface: Additive angular margin loss for deep face recognition. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019 (pp. 4690-4699).

[2] Lin K, Zhao H, Lv J, Li C, Liu X, Chen R, Zhao R. Face detection and segmentation based on improved mask R-CNN. Discrete dynamics in nature and society. 2020 May 1; 2020.

[3] Fang Z, Ren J, Marshall S, Zhao H, Wang Z, Huang K, Xiao B. Triple loss for hard face detection. Neurocomputing. 2020 Jul 20; 398:20-30.

[4] Zhao J, Han J, Shao L. Unconstrained face recognition using a set-to-set distance measure on deep learned features. IEEE Transactions on Circuits and Systems for Video Technology. 2017 May 31; 28(10):2679-89.

[5] Yan Y, Ren J, Zhao H, Sun G, Wang Z, Zheng J, Marshall S, Soraghan J. Cognitive fusion of thermal and visible imagery for effective detection and tracking of pedestrians in videos. Cognitive Computation. 2018 Feb; 10(1):94-104.

[6] Wang Z, Ren J, Zhang D, Sun M, Jiang J. A deep-learning based feature hybrid framework for spatiotemporal saliency detection inside videos. Neurocomputing. 2018 Apr 26; 287:68-83.

[7] Li S, Jiang L, Wu X, Han W, Zhao D, Wang Z. A weighted network community detection algorithm based on deep learning. Applied Mathematics and Computation. 2021 Jul 15; 401:126012.

[8] Han W, Tian Z, Zhu C, Huang Z, Jia Y, Guizani M. A Topic Representation Model for Online Social Networks Based on Hybrid Human-Artificial Intelligence. IEEE Transactions on Computational Social Systems. 2019 Dec 31.

[9] Han W, Tian Z, Huang Z, Li S, Jia Y. Topic representation model based on microblogging behavior analysis. World Wide Web. 2020 Nov; 23(6):3083-97.

[10] Verdoliva L. Media forensics and deepfakes: an overview. IEEE Journal of Selected Topics in Signal Processing. 2020 Jun 12; 14(5):910-32.

[11] Sabour S, Frosst N, Hinton GE. Dynamic routing between capsules. InProceedings of the 31st International Conference on Neural Information Processing Systems 2017 Dec 4 (pp. 3859-3869).

[12] Zhu K, Chen Y, Ghamisi P, Jia X, Benediktsson JA. Deep convolutional capsule network for hyperspectral image spectral and spectral-spatial classification. Remote Sensing. 2019 Jan; 11(3):223.

[13] Paoletti ME, Haut JM, Fernandez-Beltran R, Plaza J, Plaza A, Li J, Pla F. Capsule networks for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing. 2018 Oct 25; 57(4):2145-60.

[14] Zhu Z, Peng G, Chen Y, Gao H. A convolutional neural network based on a capsule network with strong generalization for bearing fault diagnosis. Neurocomputing. 2019 Jan 5; 323:62-75.

[15] Bregler C, Covell M, Slaney M. Video rewrite: Driving visual speech with audio. InProceedings of the 24th annual conference on Computer graphics and interactive techniques 1997 Aug 3 (pp. 353-360).

[16] Alexander O, Rogers M, Lambeth W, Chiang JY, Ma WC, Wang CC, Debevec P. The digital emily project: Achieving a photorealistic digital actor. IEEE Computer Graphics and Applications. 2010 May 27; 30(4):20-31.

[17] Dale K, Sunkavalli K, Johnson MK, Vlasic D, Matusik W, Pfister H. Video face replacement. ACM Transactions on Graphics (TOG). 2011 Dec 12; 30(6):1-0.

[18] Garrido P, Valgaerts L, Sarmadi H, Steiner I, Varanasi K, Perez P, Theobalt C. Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. In Computer graphics forum 2015 May (Vol. 34, No. 2, pp. 193-204).

[19] Thies J, Zollhöfer M, Nießner M, Valgaerts L, Stamminger M, Theobalt C. Real-time expression transfer for facial reenactment. ACM Transactions on Graphics (TOG). 2015 Oct 26; 34(6):183-1.

[20] Thies J, Zollhöfer M, Stamminger M, Theobalt C, Nießner M. Facevr: Real-time facial reenactment and eye gaze control in virtual reality. arXiv preprint arXiv:1610.03151. 2016 Oct 11.

[21] Thies J, Zollhöfer M, Theobalt C, Stamminger M, Nießner M. Headon: Real-time reenactment of human portrait videos. ACM Transactions on Graphics (TOG). 2018 Jul 30; 37(4):1-3.

[22] Kim H, Elgharib M, Zollhöfer M, Seidel HP, Beeler T, Richardt C, Theobalt C. Neural style-preserving visual dubbing. ACM Transactions on Graphics (TOG). 2019 Nov 8;

38(6):1-3.

[23] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks. Communications of the ACM. 2020 Oct 22; 63(11):139-44.

[24] Nirkin Y, Keller Y, Hassner T. Fsgan: Subject agnostic face swapping and reenactment. InProceedings of the IEEE/CVF international conference on computer vision 2019 (pp. 7184-7193).

[25] Tripathy S, Kannala J, Rahtu E. Icface: Interpretable and controllable face reenactment using gans. InProceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2020 (pp. 3385-3394).

[26] FaceSwap. www.github.com/MarekKowalski/FaceSwap. Accessed: 2021-05-10.

[27] Thies J, Zollhofer M, Stamminger M, Theobalt C, Nießner M. 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).

[28] DeepFakes. www.github.com/deepfakes/faceswap. Accessed: 2021-05-10. 1, 5

[29] Thies J, Zollhöfer M, Nießner M. Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics (TOG). 2019 Jul 12; 38(4):1-2.

[30] Fridrich J, Kodovsky J. Rich models for steganalysis of digital images. IEEE Transactions on Information Forensics and Security. 2012 May 8; 7(3):868-82.

[31] Cozzolino D, Poggi G, Verdoliva L. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security 2017 Jun 20 (pp. 159-164).

[32] Lyu S, Pan X, Zhang X. Exposing region splicing forgeries with blind local noise estimation. International journal of computer vision. 2014 Nov 1; 110(2):202-21.

[33] Popescu AC, Farid H. Exposing digital forgeries in color filter array interpolated images. IEEE Transactions on Signal Processing. 2005 Sep 19; 53(10):3948-59.

[34] Gallagher AC, Chen T. Image authentication by detecting traces of demosaicing. In2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops 2008 Jun 23 (pp. 1-8). IEEE.

[35] Dirik, Ahmet Emir, and Nasir Memon. "Image tamper detection based on demosaicing artifacts." 2009 16th IEEE International Conference on Image Processing (ICIP). IEEE, 2009.

[36] Ho JS, Au OC, Zhou J, Guo Y. Inter-channel demosaicking traces for digital image forensics. In2010 IEEE International Conference on Multimedia and Expo 2010 Jul 19 (pp. 1475-1480). IEEE.

[37] Bianchi T, Piva A. Image forgery localization via block-grained analysis of JPEG artifacts. IEEE Transactions on Information Forensics and Security. 2012 Feb 10; 7(3):1003-17.

[38] Fan Z, De Queiroz RL. Identification of bitmap compression history: JPEG detection and quantizer estimation. IEEE Transactions on Image Processing. 2003 Apr 8; 12(2):230-5.

[39] Luo W, Qu Z, Huang J, Qiu G. A novel method for detecting cropped and recompressed image block. In2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07 2007 Apr 15 (Vol. 2, pp. II-217). IEEE.

[40] Li W, Yuan Y, Yu N. Passive detection of doctored JPEG image via block artifact grid extraction. Signal Processing. 2009 Sep 1; 89(9):1821-9.

[41] Lin Z, He J, Tang X, Tang CK. Fast, automatic and fine-grained tampered JPEG image detection via DCT coefficient analysis. Pattern Recognition. 2009 Nov 1; 42(11):2492-501.

[42] Iakovidou C, Zampoglou M, Papadopoulos S, Kompatsiaris Y. Content-aware detection of JPEG grid inconsistencies for intuitive image forensics. Journal of Visual Communication and Image Representation. 2018 Jul 1; 54:155-70.

[43] Zhou P, Han X, Morariu VI, Davis LS. Two-stream neural networks for tampered face detection. In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2017 Jul 21 (pp. 1831-1839). IEEE.

[44] Zabalza J, Ren J, Zheng J, Han J, Zhao H, Li S, Marshall S. Novel two-dimensional singular spectrum analysis for effective feature extraction and data classification in hyperspectral imaging. IEEE transactions on geoscience and remote sensing. 2015 Feb 20; 53(8):4418-33.

[45] Güera D, Delp EJ. Deepfake video detection using recurrent neural networks. In 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS) 2018 Nov 27 (pp. 1-6). IEEE.

[46] Sabir E, Cheng J, Jaiswal A, AbdAlmageed W, Masi I, Natarajan P. Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI). 2019 Jan 1; 3(1):80-7.

[47] Nguyen HH, Tieu TN, Nguyen-Son HQ, Nozick V, Yamagishi J, Echizen I. Modular convolutional neural network for discriminating between computer-generated images and photographic images. InProceedings of the 13th

international conference on availability, reliability and security 2018 Aug 27 (pp. 1-10).

[48] Nguyen HH, Fang F, Yamagishi J, Echizen I. Multi-task learning for detecting and segmenting manipulated facial images and videos. arXiv preprint arXiv:1906.06876. 2019 Jun 17.

[49] Nguyen HH, Yamagishi J, Echizen I. Capsule-forensics: Using capsule networks to detect forged images and videos. InICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2019 May 12 (pp. 2307-2311). IEEE.

[50] Nguyen HH, Yamagishi J, Echizen I. Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467. 2019 Oct 28.

[51] Amerini I, Galteri L, Caldelli R, Del Bimbo A. Deepfake video detection through optical flow based cnn. InProceedings of the IEEE/CVF International Conference on Computer Vision Workshops 2019

[52] Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. Faceforensics++: Learning to detect manipulated facial images. InProceedings of the IEEE/CVF International Conference on Computer Vision 2019 (pp. 1-11).

[53] Dogonadze N, Obernosterer J, Hou J. Deep face forgery detection. arXiv preprint arXiv:2004.11804. 2020 Apr 6.

[54] Li L, Bao J, Zhang T, Yang H, Chen D, Wen F, Guo B. 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).

[55] Zhang K, Zhang Z, Li Z, Qiao Y. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters. 2016 Aug 26; 23(10):1499-503.

[56] Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A. Vggface2: A dataset for recognising faces across pose and age. In 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018) 2018 May 15 (pp. 67-74). IEEE.

[57] Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. InThirty-first AAAI conference on artificial intelligence 2017 Feb 12.

[58] Li S, Zhao D, Wu X, Tian Z, Li A, Wang Z. Functional immunization of networks based on message passing. Applied Mathematics and Computation. 2020 Feb 1; 366:124728..

[59] Rahmouni N, Nozick V, Yamagishi J, Echizen I. Distinguishing computer graphics from natural images using convolution neural networks. In2017 IEEE Workshop on Information Forensics and Security (WIFS) 2017 Dec 4 (pp. 1-6). IEEE.

[60] Bayar B, Stamm MC. A deep learning approach to universal image manipulation detection using a new convolutional layer. InProceedings of the 4th ACM workshop on information hiding and multimedia security 2016 Jun 20 (pp. 5-10).

[61] Liu H, Simonyan K, Yang Y. Darts: Differentiable architecture search. arXiv preprint arXiv:1806.09055. 2018 Jun 24.

[62] Baek JY, Yoo YS, Bae SH. Generative adversarial ensemble learning for face forensics. IEEE Access. 2020 Mar 4; 8:45421-31.

[63] Afchar D, Nozick V, Yamagishi J, Echizen I. Mesonet: a compact facial video forgery detection network. In2018 IEEE International Workshop on Information Forensics and Security (WIFS) 2018 Dec 11 (pp. 1-7). IEEE.