EXPLORING COMPLEMENTARITY OF GLOBAL AND LOCAL SPATIOTEMPORAL INFORMATION FOR FAKE FACE VIDEO DETECTION

Xiaohui Zhao, Yang Yu, Rongrong Ni and Yao Zhao*

Institute of Information Science, Beijing Jiaotong University, Beijing, China
Key Laboratory of Advanced Information Science and Network Technology, Beijing, China

ABSTRACT

The spread of fake face videos leads to severe social concerns, which promotes the development of detection methods for these videos. Existing patch-based methods focus on local regions to find forgery common clues, while ignoring the important role of the global information. In this paper, a novel spatiotemporal network is proposed which can better utilize the implicit complementary advantages of global and local information. Specifically, the spatial module consists of the global information stream and local information stream extracted from patches selected by attention layers. Then, the fusion features of these two streams are fed into the temporal module to further capture temporal clues. Besides, a regularization loss is designed to guide the selection of local information and the extraction of fusion temporal information with the reference to the global information. Extensive experiments on different datasets demonstrate the superiority of our framework.

***Index Terms*—** Fake face video detection, Global and local information fusion, Global information guidance, Spatiotemporal network

1. INTRODUCTION

With the remarkable development of deep learning, it is becoming increasingly easy to produce realistic fake face videos. However, these videos may be maliciously exploited by criminals, causing serious privacy and security issues to the society. Therefore, it is highly essential to develop reliable methods for detecting fake face videos. Currently, existing categories of fake face videos are divided into face identity swapping and face action reenactment. Face identity swap is the task of replacing the entire face of the target person in a video with the face of the source, and some popular techniques (e.g., Deepfakes [1] and FaceSwap [2]) and some tools (e.g., Deepfacelab [3] and Fakeapp [4]) are proposed to achieve face swapping. Face reenactment refers to control the motions of the target face in a video to follow the source person, and Face2Face [5] and Faceapp [6] are popular applications to implement face reenactment.

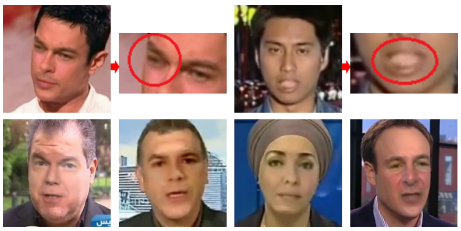

Fig. 1. Some fake faces generated by various techiniques have indiscoverable local defects (top) and global inconsistencies (bottom).

The security concerns have motivated a number of studies to detect fake face videos. Some early methods [7, 8, 9, 10, 11] employ convolutional neural networks (CNNs) and recurrent neural network (RNN) to extract features, but one limitation of these methods is that the learned feature representations may overfit to the properties of a particular database. Therefore, some patch-based detection methods are proposed to promote the improvement of generalization ability via focusing on local regions for finding tiny defects that various forgery methods left [12, 13]. However, the global information that reflects the overall consistency on detecting fake face videos is ignored in these works. Some examples are shown in Fig.1, the fake faces generated by various techiniques in the top row have indiscoverable local defects (e.g., eyes or mouth distortion), thus local information will be conducive to the detection of such fake faces. The details of the fake faces in the bottom row are realistic, but from a global perspective, all these faces have global inconsistencies (e.g., skin color unevenness and facial asymmetry). Therefore, the global information also need to be utilized on detecting fake face videos.

Motivated by above findings, in this paper, a novel spatiotemporal network is proposed to detect fake face videos, which can make full use of the implicit complementary advantages of global and local information. More specifically, we first propose a spatial module consisting of two streams: the global information stream and the local information stream extracted from patches selected by attention layers. Then, the fusion features of these two streams are fed into the

This work was supported in part by Beijing Natural Science Foundation (4222014), and National Natural Science Foundation of China (U1936212).
*Corresponding author.

proposed temporal module to further extract temporal clues. In addition, we propose a new regularization loss, termed global-stream guide (GSG) loss. This regularization aims to guide better selection of local information, and to guide fusion features to better extract temporal information with the reference to the global information. Finally, global and fusion streams are fused equally in the softmax layer to make comprehensive predictions. Moreover, extensive experiments on FF++ [14], DFDC [15] and Celeb-DF [16] datasets and comparison with state-of-the-arts demonstrate the superior generalization ability of our method.

The main contributions of this paper are summarized as follows. 1) We propose a novel spatiotemporal network that includes global and local streams to exploit the implicit complementary advantages of global and local information for comprehensive detection of fake face videos both spatially and temporally. 2) We further propose a new regularization GSG loss that utilizes the global information to guide the better selection of local information and the extraction of the temporal information from the fusion feature.

2. METHODS

2.1. Overview

The architecture of proposed method is shown in Fig.2. The framework is composed of spatial module and temporal module. The spatial module consists of the global information stream and the local information stream. Then, the fusion features of these two streams are fed into the temporal module to capture temporal detection information. In addition, we propose GSG loss to guide the selection of local information and the temporal information extraction of the fusion stream with the reference to the global information. Finally, the final output is computed by score fusion of the prediction results obtained by using global and fusion stream.

2.2. Extracting Global and Local Information

In this section, we explore the extraction of the global and local information. Specifically, we first split videos into frames and crop the faces and 12 frames are sampled from each video at a fixed interval as input. Then, we construct global and local stream based on the features extracted by ResNet-34.

Global Information. The global information of the faces is extracted using an average pooling layer. As the average pooling layer calculate the average value of the feature maps directly, each region of the images is considered equally. By this way, we acquire the global feature of the t th frames ${F_t}^{\mathrm{G}}$:

$$ {F_t}^{\mathrm{G}} = \frac{1}{h \times w} \sum_{i=1}^{h \times w} v_{t,i} \tag{1} $$

where $v_{t,i}$ is the i th vector of the activation tensor obtained from the t th frame, h and w is the height and the width of feature maps through the backbone.

Local information. As for the capture of the local information, we particularly extract local features from the feature maps rather than from original RGB inputs. We extract fixed size patches in a sliding fashion over entire images. Then we introduce attention mechanism to select patches from the split 16 patches by self-learning to focus on critical details of each image. The attention network in this paper includes two fully connected layers with 128 neurons and one neuron, respectively. The scalar values obtained by the learning of the attention layers are used to represent the importance of the patches. After applying a softmax operation to normalize the scalar values as weights, we select the top N most discriminative patches based on the weights. To aggregate the patches, we take a simple average of the activation values in every patch and pool the mean into the same size as the global feature. Therefore, we acquire the local feature ${F_t}^{\mathrm{L}}$:

$$ {F_t}^{\mathrm{L}} = \mathrm{pool}(\frac{1}{N} \sum_{j=1}^{N} p_{t,j}) \tag{2} $$

where $p_{t,j}$ is the vector of the j th patch of the t th frame.

After obtaining the global information and local information, we fuse the two features by concatenating them to get the composite features F_t:

$$ F_t = \mathrm{concatenate}({F_t}^{\mathrm{G}}, {F_t}^{\mathrm{L}}) \tag{3} $$

2.3. Capturing Temporal Information

Since the forgery process is operated frame by frame, the inter-frame correlation is destroyed, illustrating that the temporal discontinuity needs to be captured in fake face video detection. Therefore, we design a temporal module to capture the dynamic inconsistency clues. Previous methods usually extract temporal clues via 3DCNN [17], RNN [18] or LSTM [19], which have a high complexity. To solve this problem, in our framework, SRU unit [20] is employed instead of RNN or LSTM unit. In the SRU unit, the update of gate states no longer depends on the previous hidden states. It simplifies the calculation by dropping the previous hidden states, which improves the parallel capability without performance degradation. Specifically, the proposed temporal module consists of three layers of SRU unit with a fully connected layer and an average pooling layer behind, as showed in Fig.2. Then the softmax layer is used to obtain the prediction results of the fusion stream. In addition, the predictions of frames belonging to a video are averaged as the final prediction of the video.

2.4. Global-Stream Guide Loss

In this section, we propose a new regularization loss, termed global-stream guide (GSG) loss. This loss aims to use the global information to guide the temporal module to extract more sufficient and reliable temporal information from the fusion feature and to promote the selection of local information. Specifically, L_{GSG} is defined as:

$$ L_{\mathrm{GSG}} = \max(0, P_{\mathrm{G}} - P_{\mathrm{F}} + \mathrm{margin}) \tag{4} $$

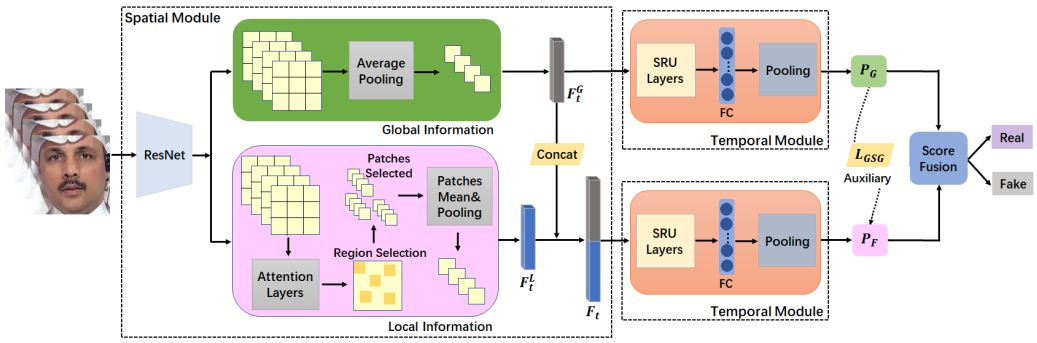

Fig. 2. The architecture of proposed method.

where P_{G} and P_{F} are the prediction probabilities obtained by inputting global features or fusion features into the temporal module, respectively. It can be seen that the smaller the L_{GSG} is, the larger the P_{F} is. In other words, the goal of minimizing L_{GSG} is to make the prediction of the stream using fusion features more accurate, leading to the better extraction of spatiotemporal information and the better selection of local information. Concretely, the margin value is set to 0.1. Therefore, the total loss L of the model is formulated as:

$$L = L_{\mathrm{CE}} + L_{\mathrm{GSG}} \tag{5}$$

where L_{CE} is the cross-entropy loss.

Finally, the final detection result of our framework is predicted via the score fusion of the prediction values between the global stream and fusion stream, and the experimental setup in this paper is to average these two stream results. By combining these two representations, the reliability of the video-level representation is further enhanced.

3. EXPERIMENTS

3.1. Datasets and Implement Details

We evaluate the proposed method on three popular fake face video datasets: FaceForensics++(FF++) [14], DFDC [15], and Celeb-DF [16]. FF++ dataset consists of 5000 videos including four types of fake videos (DeepFakes, Face2Face, FaceSwap and NeuralTextures) and corresponding real videos. For DFDC dataset, the various generation technologies and the excellent forgery quality of fake videos lead its detection difficulties, and 1000 fake videos and 1000 real videos of DFDC dataset are utilized. Celeb-DF consists of more than 5000 deepfake videos, and the real videos are gathered from social media.

In the experiments, we use MTCNN [21] to detect the faces and adopt a conservative crop to enlarge the face region by a factor of 1.3 around the center of the tracked face, following the setting in [14], and then the extracted faces are aligned to size of 224×224. The size of patches extracted from feature maps is 2×2. The models are trained in an end-to-end fashion and we use the SGD optimizer to optimize the network. The initial learning rate is set as 10^{-2} or 10^{-3}.

Table 1. Intra-dataset evaluation on FF++, DFDC and Celeb-DF (Acc %).

Model	FF++		DFDC	Celeb-DF
	LQ	HQ		
MesoNet [7]	68.13	81.80	89.33	86.93
Xception [8]	81.73	91.67	94.83	94.59
Face x-ray [23]	69.47	86.53	96.83	97.97
LipForensics [24]	81.67	93.93	97.33	98.56
MaDD [12]	82.80	94.00	97.50	98.11
Patch [13]	81.27	94.07	97.17	97.32
RNNs [10]	81.93	93.80	97.33	97.13
Two-branch [11]	82.20	93.87	97.50	98.20
Ours	**83.53**	**95.33**	**97.67**	**99.18**

Table 2. Cross-dataset evaluation on Celeb-DF and DFDC by training on FF++ (Acc %).

Model	Celeb-DF	DFDC
MesoNet [7]	57.80	53.80
Xception [8]	69.67	63.35
Face x-ray [23]	79.57	62.90
LipForensics [24]	81.58	67.10
MaDD [12]	72.18	64.70
Patch [13]	79.49	65.65
RNNs [10]	74.80	62.75
Two-branch [11]	79.32	66.50
FTCN [25]	83.58	68.80
Ours	**88.83**	**69.25**

3.2. Comparison Experiments

In this section, we compare our framework with current state-of-the-art detection methods. All results are video-level obtained by merging the scores of frames and all methods use the same number of frames.

Intra-dataset Evaluation. Firstly, we evaluate the intra-dataset performance on FF++, DFDC and Celeb-DF datasets. For fair comparisons, all methods are used in the same experimental settings. Each dataset is split into training set and test set at a ratio of 7:3. The comparative results are shown in Table.1. It can be found that the proposed model outperforms state-of-the-art methods on these three datasets. Especially, the accuracy of our approach is up to approximately 1% or 2% higher than the patch-based methods [12, 13] and temporal-based methods [10, 11] on FF++ and Celeb-DF dataset.

Cross-dataset Evaluation. We further conduct cross-dataset experiments to evaluate the generalization capability of the proposed method. Specifically, we train the models

Table 3. Ablation study of the parts of our framework on Celeb-DF (Acc %).

Model	Celeb-DF
Global information	96.65
Local information	92.51
w/o feature fusion	98.47
w/o temporal module	98.12
w/o GSG loss	97.03
w/o score fusion	98.11
RGB patches	98.57
Final model	**99.18**

Table 4. Ablation results on Celeb-DF dataset with different number of patches selected from each frame (Acc %).

Model	Celeb-DF	Model	Celeb-DF
Number=4	98.85	Number=7	98.68
Number=5	**99.18**	Number=8	98.52
Number=6	99.02	Number=9	98.35

on FF++ dataset and test on Celeb-DF and DFDC dataset. Table.2 shows the comparative results. It is observed that although the cross-dataset performance is decreased than that of intra-dataset, the result of our method is also better than other methods, which proves that our method has better generalization ability.

3.3. Ablation Study

In order to verify the effectiveness of each component of our framework, we perform the ablation study on the Celeb-DF dataset. The results are shown in Table.3.

Firstly, in order to analyze global and local information, we utilize global stream and local stream to detect Celeb-DF dataset, respectively, and we further test the framework without (w/o) feature fusion. The results are shown in the first three lines of Table.3, respectively. From the results, the performance of using dual-stream as input is better than using single stream as input, but the performance of the model still drops if there is no feature fusion operation, which verifies the effectiveness of the feature fusion and also reveals the potential complementary advantages of global information and local information. Then, we evaluate the performance of our framework without temporal module, and the result is shown in the fourth row of Table.3. It can be seen that the proposed temporal module benifts the extraction of the temporal discontinuities. Next, we evaluate the framework without GSG loss and score fusion and show the results in Table.3. The result also proves the effectiveness of these two modules.

To further study the effectiveness of using feature maps for local feature extraction as mentioned in Sec.2.2, we also implement a framework for cropping patches at the predicted positions from the original RGB inputs. The result shows in Table.3 illustrates that extracting local features from the feature maps can help to obtain more discriminative information.

As mentioned in Sec.2.2, we select fixed size patches from the images after learning the weights of all patches over entire images by the attention layers. The selected patches can focus on critical details of frames. We determine the number of patches through experiments and show the results in Table.4. According to the results, the model achieves the best performance when the number of patches is 5. In addition, the selection of the number of frames is also experimentally verified, as shown in Table.5. There is a slight increase in performance from sampling 12 to 16 frames, so we choose 12 frames to save computational cost.

Table 5. Ablation study on Celeb-DF dataset with different numbers of frames sampled from each video (Acc %).

8	9	10	11	**12**	13	14	15	16
98.91	99.02	99.09	99.14	**99.18**	99.18	99.19	99.20	99.20

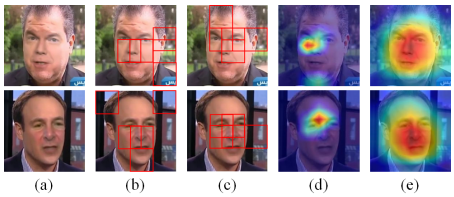

Fig. 3. Visualization of the selected patches and the heatmaps extracted by gradcam.

3.4. Examples visualization

To further illustrate the complementary advantages of global and local information, we visualize the selected patches and the heatmaps extracted by gradcam [22] of fake faces, as shown in Fig.3. Specifically, Fig.3 (a) are the fake faces with realistic details but with global skin color inconsistency (top) and facial asymmetry (bottom). Fig.3 (b) and Fig.3 (c) show the patches selected by our model using only local information and the model with the guidance of global information, respectively. It can be seen that the global information helps to meaningfully select local patches to regions with inconsistent skin colors and facial asymmetry. In addition, Fig.3 (d) are the heatmaps of a patch-based method [12], indicating that the key information of artifacts are not be captured. While Fig.3 (e) is the heatmaps of our method after adding global information, indicating that the complementarity of two streams information helps our framework to consider inconsistent fake clues and obtain correct detection results.

4. CONCLUSION

In this paper, a novel spatiotemporal network is proposed which can make full use of the implicit complementary advantages of global and local information. We first extract the spatial information of the video frames from the global and local perspectives. Then the fusion features of these two branches are fed into the temporal module to further extract temporal clues. Moreover, a new regularization loss termed global-stream guide (GSG) loss is designed to guide the selection of local information and the extraction of temporal information of the fusion stream. Extensive experiments on three popular datasets and comparison with state-of-the-arts demonstrate the superior performance of our framework.

5. REFERENCES

[1] Deepfakes.https://github.com/deepfakes/faceswap

[2] FaceSwap.https://github.com/deepfakes/faceswap-playground

[3] Deepfacelab.https://github.com/iperov/DeepFaceLab/. Accessed: 2019-08-20. 1

[4] Fakeapp.https://www.fakeapp.com/. Accessed:2019-07-25. 2

[5] Thies, Justus. "Face2Face: Real-Time Facial Reenactment." Information Technology, vol. 61, 2019, pp. 143–146.

[6] Faceapp, https://github.com/lolPants/faceapp.js

[7] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen, "Mesonet: a compact facial video forgery detection network," in 2018 IEEE International Work shop on Information Forensics and Security (WIFS), 2018, pp. 1–7.

[8] Francois Chollet, "Xception: Deep learning with depthwise separable convolutions," in CVPR, 2017.

[9] Nguyen, Huy H., et al. "Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos." ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2307–2311.

[10] Sabir, Ekraam, et al. "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 80–87.

[11] Masi, Iacopo, et al. "Two-Branch Recurrent Network for Isolating Deepfakes in Videos." ECCV (7), 2020, pp. 667–684.

[12] Zhao, Hanqing, et al. "Multi-Attentional Deepfake Detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2185–2194.

[13] Lucy Chai, David Bau, Ser−Nam Lim, and Phillip Isola, "What makes fake images detectable? understanding properties that generalize," in ECCV, 2020.

[14] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner, "Faceforensics++: Learning to detect manipulated facial images," in ICCV, 2019.

[15] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge dataset. arXiv preprint arXiv:2006.07397, 2020.

[16] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb−df: A large−scale challenging dataset for deep fake forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3207–3216, 2020.

[17] Zhang, Daichi, et al. "Detecting Deepfake Videos with Temporal Dropout 3DCNN." Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, vol. 2, 2021, pp. 1288–1294.

[18] Guera, David, and Edward J. Delp. "Deepfake Video Detection Using Recurrent Neural Networks." 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), 2018, pp. 1–6.

[19] Amerini, Irene, and Roberto Caldelli. "Exploiting Prediction Error Inconsistencies through LSTM-Based Classifiers to Detect Deepfake Videos." Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security, 2020, pp. 97–102.

[20] Lei, Tao, et al. "Simple Recurrent Units for Highly Parallelizable Recurrence." Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 4470–4481.

[21] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters 23, 10 (2016), 1499–1503.

[22] Selvaraju, Ramprasaath R., et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.

[23] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo, "Face x-ray for more general face forgery detection," CVPR, June 2020.

[24] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," CVPR, June 2021.

[25] Y. Zheng, et al. "Exploring Temporal Coherence for More General Video Face Forgery Detection." ICCV 2021.