2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506371

FACE FORGERY DETECTION BASED ON SEGMENTATION NETWORK

Yingbin Zhou [1] *, *Anwei Luo* [2] *, *Xiangui Kang* [2] †, *Siwei Lyu* [3]

[1] School of Systems Science and Engineering, Sun Yat-Sen University, Guangzhou, China 510006
[2] School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, China 510006
[3] Department of Computer Science and Engineering, State University of New York Buffalo, NY14260

ABSTRACT

Recent progress in facial manipulation technologies have made it hard to distinguish the sophisticated face swapped images/videos. Due to the diversity of generation software and data sources, it is extremely challenging to devise an efficient generality framework. Instead of regarding the detection process as a vanilla binary classification task, we proposed a detection framework based on pixel-level classification. Considering that the acquisition of real pixel-level ground-truth is somehow expensive or even impractical, we proposed a pseudo ground-truth generation pipeline with prior knowledge of facial manipulation. Besides, we added a new module into the neural network to capture frequency clues, while the ablation experiment verified the effectiveness of this module. The experimental results on several public datasets demonstrated that our proposed framework is effective and superior to other existing similar detection networks.

Index Terms**—**face swapped images/videos, pixel-level classification, pseudo ground-truth generation, frequency clues

1. INTRODUCTION

With the significant progress in deep learning and the easy availability of a large amount of training data in the age of social networks, manipulated multimedia becomes harder and harder to distinguish by human eyes, which dramatically reduces the credibility of digital media. One popular type of manipulated method called deepfake (or faceswap) has drawn tremendous attention in recent years. As shown in Fig. 1, the facial identity of person A is swapped by another person B through deep learning models, such as Generative Adversarial Networks (GAN) [1] or Auto-Encoders (AE) [2], while the original expression and posture are preserved. Since facial features are essential for identity recognition, any rumormonger's fake multimedia would raise social security issues even cause public panic. Therefore, it is vital to develop relative countermeasures to prevent the spread of misleading information.

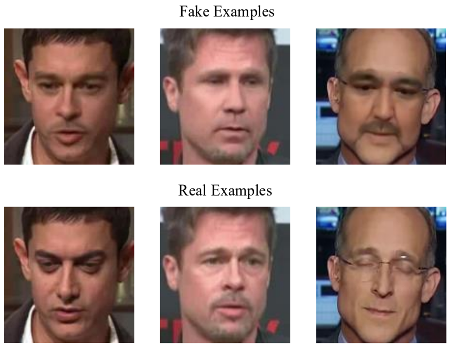

Fig. 1. Examples of real images and facial forged images.

Various organizations and individuals have contributed to fight against face forgery [3][4]. For the purpose of facilitating researchers to develop advanced detection methods, several large public datasets were released in recent years, in which Faceforensics++ dataset [5] and Celeb_DF dataset [6] are the most widely used dataset in manipulated detection community. In [7], tremendous genuine and forged videos were collected from the internet, which aims to develop and test the effectiveness of detectors against real-world scenario.

Regarding detection methods, in the early period, researchers tried some hand-crafted features to discover forgery patterns, such as steganalysis features [8] and color components analysis [9], but these methods may fail when forged images/videos were undergone post-processing (i.e., JPEG compression or H. 264 coding). Current state-of-the-art methods catch the artifacts by incorporating deep learning technologies. In [5], several Convolutional Neural Networks (CNNs) were used to evaluate their effectiveness on four standard forgery methods (includes DeepFakse [10], FaceSwap[11], Face2Face [12], and NeuralTextures [13]). Among these methods, MesoNet [14] and MISLNet [15] showed high accuracy on the raw facial forged videos, while Xception [16] exhibited the best performance and became a strong benchmark for future study. Nguyen et al. built a general detection framework based on Capsule neural network and the feasibility was demonstrated by verifying a wide range of forged attacks [17]. Li et al. observed the face

[*] Equal contribution.
[†] Corresponding author (e-mail: isskxg@mail.sysu.edu.cn).

This work was supported by NSFC (Grant nos. 61772571, 62072484) and Chinese national key research and development project 2019QY2203.

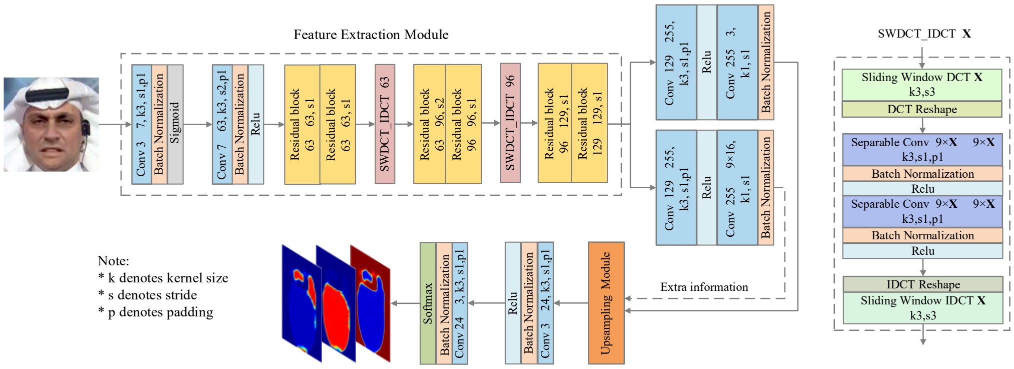

Fig. 2. Diagram of the proposed neural network architecture.

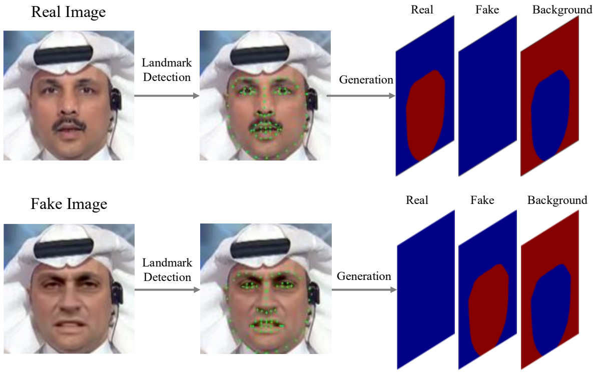

Fig. 3. The pipeline of pseudo ground-truth generation.

warping artifacts in deepfakes and constructed an extra dataset to magnify these clues between real and fake images [18]. A similar idea could be found in [19], where Face X-ray was devised to detect the trace of the modification boundary of facial forged images. Note that some of the methods mentioned above occur dramatic performance decline after video compression, so the practical application is limited. Besides, due to the lack of time continuity in deepfake videos, time continuity analysis [20][21][22] and biological signal detection [23] were also utilized to capture the discrepancy between genuine and manipulated videos.

The contributions of this paper are as follows: 1) we designed a deep neural network (as shown in Fig. 2) by incorporating the idea of image segmentation and took advantage of the prior knowledge that identity swap always happened within facial area to generate pseudo ground-truth for network training; 2) we devised a new module called Sliding Window DCT_IDCT module to fuse the frequency information into the neural network; 3) Compared with other state-of-the-art methods, our proposed face forgery detection framework exhibits the top performance on several public datasets and it is robust to compression operation.

2. METHOD

Our proposed face forgery detection framework mainly includes three parts: pseudo ground-truth generation, deep learning model and prediction criterion. We first use the prior knowledge to generate the pixel-level pseudo ground truth, then the deep learning model is trained by the supervision of pseudo ground-truth. Finally, the classification result is obtained by using an ensemble prediction criterion. The detail of each part is described in the following sections.

2.1. Pseudo Ground-truth Generation

Unlike other face forgery detection methods regarding facial manipulation forensics as a vanilla binary classification problem, we consider face forgery detection as an ensemble result by using a segmentation network. Since it is expensive or even impossible to obtain the real pixel-level ground-truth, an alternative way is to generate pseudo ground-truth (PGT) by the prior knowledge that the pixels within the facial area of a forged image tend to be modified with a high probability *p*. The pipeline of pseudo ground-truth generation is shown in Fig. 3. Given a suspected facial image, we first extract 68-points landmark with Dlib package [24] and generate a facial region mask. Considering that the post-processing operation (e.g., Possion blending) may enlarge the modified region, in order to have a better coverage for the PGT, we dilate the facial mask to contain a few background regions. In the following step, we create three PGTs (real facial PGT G^r, fake facial PGT G^f and background PGT G^b) to represent the probability of the corresponding class. In practically, for a real training sample, we assign pixel-level label value of [*p, (1-p)/2, (1-p)/2*] to [G^r, G^f, G^b] within the facial region mask, while we assign [*(1-p)/2, p, (1-p)/2*] to [G^r, G^f, G^b] for a fake training sample. Regarding to background region, [0, 0, 1] is assigned to [G^r, G^f, G^b] for both real and fake training samples. It is worthy to note that we devise a three-PGTs scheme rather than a binary (real or fake) supervision fashion to guide the network to pay more attention to the probably manipulated facial region during the training process.

2.2. Neural Network Architecture

The proposed detection framework is based on a light and shallow neural network, its architecture is illustrated in Fig.

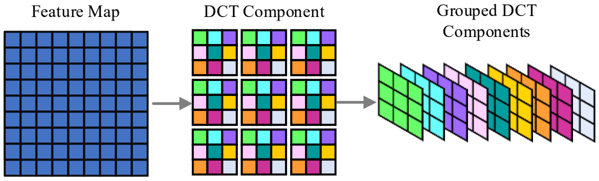

Fig. 4. The process of sliding window DCT.

2. The feature extraction module contains two common convolution layers and six residual blocks, while the SWDCT_IDCT module is devised and inserted into the neural network to enhance feature extraction ability. In this module, a Sliding Window DCT (SWDCT) is used to distill local frequency information from input feature maps. SWDCT is implemented by assigning two-dimensional DCT coefficients to the convolution kernel. After DCT transformation, as shown in Fig. 4, the frequency components of each feature map are grouped separately through reshaping operation according to different frequency components. Then two separable convolution layers are utilized to process the frequency components and the frequency information is converted back to the original domain through Sliding Window IDCT (SWIDCT). Using the frequency-aware information to capture the forgery clues have been investigated in prior works [25][26], where frequency-aware information only happened in the beginning of proposed networks. In our work, frequency information is collected and optimized from the middle level feature maps, while the relationship of the spatial information is retained after SWIDCT. Furthermore, we adopt the upsampling module described in [27] to process the low-resolution feature maps (i.e., three channel feature maps with size 64 × 64 in our experiment). Different to bilinear upsampling operation, each pixel of the high-resolution outputs is taken to be a convex combination of its 9 coarse resolution neighbors via using weights predicted by the network (i.e., the extra information shown in Fig. 2). The final predicted outputs represent the pixel-level probability to real, fake or background.

2.3. Loss Function and Prediction Criterion

The proposed framework is trained by using cross-entropy loss function, and the definition is as follow:

$$\mathrm{Loss} = \frac{1}{\mathrm{N}}\sum\sum_{\mathrm{C}} -\mathrm{G_C}\log(\mathrm{O_C}) \qquad (1)$$

where N is the number of the pixels, $\mathrm{G_c}$ and $\mathrm{O_c}$ represent the assigned labels and predicted output probabilities corresponding to real face class, fake face class and background class. For simplicity, we omit the location index.

The forgery operation is mainly focused on the facial region, so we calculate the category scores by summing up the corresponding probability within the facial mask region. Generally, since the proposed method can be considered as an ensemble framework of pixel-level supervision learning, we do not need each pixel position to predict exactly. Suppose most pixels could get a satisfactory prediction score, in that case, the final result calculated by the ensemble criterion will be accurate, which means that our proposed face forgery framework is less prone to overfitting, while the robustness and generation ability when facing different face forgery situations are guaranteed.

3. EXPERIMENT

In this section, we conduct a systematic analysis and evaluate the proposed face forgery detection framework on several large public datasets.

3.1. Experimental Setting

Dataset The performance of our framework was evaluated by several public datasets: FaceForensics++ dataset, Celbe_DF dataset and Wild_DF dataset. FF++ dataset contains 1000 real videos and 4000 fake videos created by four forgery methods. Note that we only conducted experiments on DeepFake (DF) method and FaceSwap (FS) method. Celeb_DF dataset contains 590 real videos and 5639 high-quality fake videos generated by using an improved synthesis process, while the low visual distortion in this dataset makes the detection become harder. Wild_DF dataset contains 3805 real face sequences and 3509 fake face sequences collected from the internet, it is the most challenging dataset for our experiments. The data splitting strategy is strictly followed the original criteria in each dataset. We randomly selected 50 frames in each video or sequence to build the frame-level experiments.

Parameter setting In the experiments, all the face images were detected with MTCNN [28] then resized to 256 × 256. All experiments were performed on PyTorch package on Nvidia 3090 GPUs. The model was trained from scratch by using Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and the weight decay 1e-8. The initial learning rate was set to 2e-3 and decreased with reduce_factor = 0.1 after every 5 epochs. Since the PGTs are devised with prior knowledge and we do not know the real ground-truth, p was set to 0.9 to introduce uncertainty. We applied the accuracy score (***Acc***) and Area Under the Recevier Operating Characteristic Curve (***AUC***) as the evaluation metrics. Considering the heavy class-imbalance in Celeb_DF dataset, we only report its ***AUC*** results.

3.2. Performance of the Proposed Framework

We compared our proposed method with several state-of-the-art methods in the same condition. Table I summarized the test performances of their optimized models. The performance of our proposed method on FF++ dataset and Celeb_DF dataset is superior to other methods, and the classification accuracy almost achieves 100%. The videos in Wild_DF dataset is extremely challenging for the detectors

Table I. Performance Comparisons on Different Dataset

Method	FF++_DF		FF++_FS		Celeb_DF		Wild_DF	
	Acc	***AUC***	***Acc***	***AUC***	***Acc***	***AUC***	***Acc***	***AUC***
Xception [16]	97.85%	0.9982	96.92%	0.9949	-	0.9977	76.26%	0.8414
MesoNet [14]	93.85%	0.9886	89.98%	0.9717	-	0.9377	73.95%	0.8321
Capsule [17]	98.54%	0.9986	98.71%	0.9989	-	0.9967	78.65%	0.8631
MISLNet [15]	96.06%	0.9938	91.36%	0.9722	-	0.9745	73.70%	0.8145
Proposed	**99.16%**	**0.9994**	**99.76%**	**1.0000**	-	**0.9978**	**79.87%**	**0.8782**

Table II. Performance on Low Quality Version FF++ Dataset

Method	FF++_DF(C40)		FF++_FS(C40)	
	Acc	***AUC***	***Acc***	***AUC***
Xception [16]	95.01%	0.9903	91.23%	0.9724
MesoNet [14]	88.26%	0.9524	84.38%	0.9232
Capsule [17]	94.68%	0.9853	91.34%	0.9650
MISLNet [15]	90.48%	0.9715	87.90%	0.9496
Proposed	**96.10%**	**0.9917**	**96.09%**	**0.9933**

due to the complexity in real scenario, the results of ***Acc***=79.87% and ***AUC***=0.8782 obtained by our proposed detector are dominant. Compared to the method [16], which is the strong baseline in face forgery detection field, our proposed deep learning model is a relatively small network (3.8 million parameters). The improvements of 3.61% in ***Acc*** and 0.0368 in ***AUC*** indicate that the pixel-level training fashion in our proposed method helps to guide the network capture artifacts in the forged images. Besides, considering that digital products often need to be compressed before uploading to the internet, we conducted a comparison experiment on low quality version FF++ dataset (i.e., compressed videos by H. 264 codec with constant rate quantization of 40). Since the compression operation would erase the forensics clues, it can be seen from Table II that the detection performance under different methods is inevitably reduced. Compared to the Xception network, the performance of our proposed method is particularly impressive when fighting against compression, leading to the improvements of 1.09% on FF++_DF and 4.86% on FF++_FS when using ***Acc*** as the metric. The SWDCT_IDCT module may contributes to these improvements, since the information contained in some frequency bands may not sensitive to compression operation, the frequency artifacts would be preserved even in a heavy compression condition, so utilize these artifacts can promotes the detection performance.

3.3. Ablation Study

In this section, we conducted an ablation experiment to verify the effectiveness of the proposed SWDCT_IDCT module. To get a relatively fair comparison, we replaced the SWDCT_IDCT module by using naive residual block and guaranteed that the number of layers in both networks is the same. The results on low quality version FF++_DF dataset are shown in Table III. The proposed method got the improvements of 1.91% in ***Acc*** and 0.0048 in ***AUC*** when the network is equipped with SWDCT_IDCT module. We summarized two advantages for using this module: 1) Current facial forgery models do not consider the frequency information in the generation process, although the generation quality is very high in the spatial domain, the neglect of information consistency in the frequency domain makes an efficient detection performance by using the frequency artifacts; 2) The information in frequency domain is optimized with the separable convolutions, then converted back to the spatial domain. SWDCT_IDCT module not only improves the attention to the artifacts by using the frequency information but also preserves the position relationship in the spatial domain.

Table III. Effetiveness of SWDCT_IDCT Module

Method	***Acc***	***AUC***
Rsidual Block	94.19%	0.9869
SWDCT_IDCT	**96.10%**	**0.9917**

4. CONCLUSIONS

In this paper, we proposed a novel face forgery detection perspective to promote the progress in the multimedia forensic field. Utilizing the prior knowledge of facial manipulation, we proposed the idea of PGT to guide network training and designed an ensemble mechanism to improve the robustness and generation ability, which is realized by devising a three-classification segmentation network and a calculation criterion. We also incorporated frequency information to enhance the artifact capture ability by using the SWDCT_IDCT module, while the original spatial location information is maintained. The experiments conducted on several public datasets show that our approach outperforms other similar methods, which demonstrates the effectiveness of the proposed face forgery detection framework. In the future work, we will investigate a more sophisticated pseudo ground-truth generation mechanism and also explore the interpretability of the proposed detection framework.

5. REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu,D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. of the 28th Conference on Neural Information Processing Systems* (NIPS), 2014, pp. 2672‑2680.

[2] I. Goodfellow, Y. Bengio, A. Courville and Y. Bengio, "Deep Learning," *Cambridge: MIT Press*, 2016.

[3] https://ai.googleblog.com/2019/09/contributing-data-to-deepfakedetection.html.

[4] https://deepfakedetectionchallenge.ai/dataset.

[5] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Niener, "Faceforensics++: Learning to detect manipulated facial images," in *Proc. of the IEEE International Conference on Computer Vision* (ICCV), 2019: 1-11.

[6] Y. Li, X. Yang, P. Sun and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 2020: 3207-3216.

[7] B. Zi, M. Chang, J. Chen, X. Ma and Y. Jiang, "WildDeepfake: a challenging real-world dataset fordeepfake detection," in *Proc. of the 28th ACM International Conference on Multimedia* (ACM MM), 2020: 2382-2390.

[8] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 868–882, 2012.

[9] H. Li, B. Li, S. Tan and J. Huang, "Detection of deep network generated images using disparities in color components," *arXiv preprint arXiv*:1808.07276, 2018.

[10] https://github.com/deepfakes/faceswap.

[11] https://github.com/MarekKowalski/FaceSwap.

[12] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proc. of the IEEE conference on Computer Vision and Pattern Recognition* (CVPR), 2016, pp. 2387‑2395.

[13] J. Thies, M. Zollhofer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Transactions on Graphics*, vol. 38, no. 4, pp. 1‑12, 2019.

[14] D. Afchar, V. Nozick, J. Yamagishi and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *Proc. of 2018 IEEE International Workshop on Information Forensics and Security* (WIFS), 2018, pp. 1-7.

[15] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proc. of the 4th ACM Workshop on Information Hiding and Multimedia Security* (IH&&MM), 2016, pp. 5-10.

[16] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. of the IEEE conference on computer vision and pattern recognition* (CVPR), 2017, pp. 1251-1258.

[17] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *Proc. of 44th IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), 2019, pp. 2307-2311.

[18] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," *arXiv preprint arXiv:1811.00656*, 2018.

[19] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-Ray for more general face forgery fetection," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), IEEE, 2020, pp. 5001-5010.

[20] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. of the15th IEEE International Conference on Advanced Video and Signal Based Surveillance* (AVSS), 2018, pp. 1-6.

[21] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces* (GUI), 2019.

[22] I. Amerini, L. Galteri, R. Caldelli and B. Del, "Deepfake video detection through optical flow based cnn," in *Proc. of the IEEE/CVF International Conference on Computer Vision Workshops* (CVPRW). 2019.

[23] U. A. Ciftci, I. Demir, and L. Yin, "Fakecatcher: Detection of synthetic portrait videos using biological signals," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[24] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755-1758, 2009.

[25] K. Xu, M. Qin, F. Sun, Y. Wang, Y. K. Chen and F. Ren, "Learning in the frequency domain," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 2020, pp. 1740-1749.

[26] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, "Thinking in frequency: Face forgery detection by mining frequency-aware clues," In *Proc. of European Conference on Computer Vision* (ECCV), 2020, pp. 86-103.

[27] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in Proc. of *European Conference on Computer Vision* (ECCV), 2000, pp. 402-419.

[28] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499-1503, 2016.