A deep learning approach to face swap detection

Svetlana Volkova, Alexey Bogdanov

***Abstract*—We propose the method for detecting an incident at face authentication when an imposter falsifies the client's face using a digital technique named Face Swap to cheat the system. The method is based on a convolutional neural network to get facial features and classify them. The proposed method can work with faces obtained with low quality and heavy lighting conditions. It is confirmed by experiments on a big test dataset. Experiments show that the accuracy reaches values over 98% for low-quality images and over 99% for high-quality images. Classification results are congruent to the best results shown by the other known methods tested on the same test dataset. The proposed method can be applied to improve the quality of face authentication systems.**

***Keywords*—Face swap, face forgery detection, DeepFake detection, face replacement, image forensics, deep learning.**

I. Introduction

In recent years there have been a lot of techniques for face manipulation.

The face manipulation technique (or DeepFake technique) can be used in a different area, where people's photos or videos are used. For example, after the tragic death of Paul William Walker, creators of The Fast and the Furious used deepfakes to finish shooting [1].

However, there are negative aspects of using these techniques. The consequences of the hostile use of such technology could harm by providing misinformation or fake news.

For example, In May 2018 a fake video with Donald Trump (who at that time was fulfilling his first term of presidency of the USA) appeared on the Internet. In this video, a fake "Trump" appealed to the Kingdom of Belgium to follow America's lead and exit the Paris climate agreement (Fig. 1) [2].

In addition, with the electronic document management development, using digital tampering techniques can reduce the security of the face authentication system, because an imposter can falsify the client's face. This is complicated by the fact that face manipulation techniques can be used without any special equipment, and they are now capable of running on mobile phones.

For example, In January 2021 the People's Prosecution of Shanghai City charged the two offenders, with fraud with a facial recognition system. The perpetrators have been cheated on verification of the identity system of the Internal Revenue Service since 2018. For the circumvents of the system, the offenders used high-quality photographs of other people and their personal data. In the end, they used DeepFake applications to forge data. The photos thus processed were sent to the biometric system. In this scheme, the scammers were able to earn about $ 77 million [3].

Fig. 1 - Frame of the fake video featuring D. Trump: an example of the use of deepfake technique.

There are a lot of applications for this kind of manipulation and most of them can be downloaded for free in the Play Market / App Store. At the same time, the quality of face transfer is quite good, and a human being cannot define whether the manipulation was done or not. It shows that the development of automated face swap detection methods is important.

The face manipulation techniques can be separated into several categories. The first category is GAN-based methods for face synthesis [4, 5, 6]. The second ones try to manipulate facial expressions. The most popular technique for facial expression manipulation is Face2Face [7], which can transfer facial expression from one person to another in real-time. The third category named facial identity manipulation is the most dangerous for biometric systems. This category is known as a face swap [8, 9]

In this research, we proposed a method for increasing the security of face authentication systems by detecting swapped faces on the image. This paper is organized as follows. Section II summarizes the deep learning approaches to swapped face detection. Section III presents a proposed

Manuscript received June 30, 2021.

S. S. Volkova. Candidate of science (Ph.D.), Vologda State University, Applied Math Department, Vologda, Russia (corresponding author to provide phone: +7-905-298-74-17; e-mail: malysheva.svetlana.s@gmail.com).

A. S. Bogdanov was with the Vologda State University, Applied Math Department, Vologda, Russia.

FaceSwap detection method. Section IV describes an experiment to evaluate the performance of the proposed method and to compare it with other detectors. Finally, we conclude the whole paper.

II. Related works

The subject of face forgery detection is actively studied [10, 11].

Deepfake detection methods can fall into two categories. Methods in the first category focus on the detection artifacts, which are generated by DeepFake techniques. There are a few types of artifacts that are analyzed: blending analysis [12, 13, 14, 15], environment analysis [16, 17], behavior analysis [18, 19], coherence analysis [17, 20]. The second category focuses on training generic classifiers, instead of focusing on specific artifacts.

In this section, we will discuss various non-artifact-specific approaches to deepfake detection. These methods fall into two categories: global classification and anomaly detection.

Anomaly detection methods are the deep learning approaches that are trained on the normal data and then detect outliners during deployment.

Wang et al. [21] proposed an approach, named FakeSpotter. It is an approach for fake face detection based on neutron coverage. Here, they conjecture that monitoring neutron behavior can serve as an asset in detecting fake faces since layer-by-layer neutron activation patterns may capture more subtle features.

In the paper [22] authors proposed the application of the attribution-based-confidence (ABC) metric for detecting deep fake videos. The ABC metric does not require access to the training data or training the calibration model on the validation data.

The anomaly detection method, which is proposed in [23], consists of a preprocessing step where the content of the image is suppressed, and the anomaly locations and anomaly strengths are extracted. The classification is then done by a simple classifier.

Training a global classifier is another category of non-artifact-specific approaches. Various authors demonstrated different neural network architectures as applied to deep fake detection task. In [24] the authors trained Xceprion Net for deep fake detection. Paper [25] follows a deep learning approach and presents two networks, both with a low number of layers to focus on the mesoscopic properties of images.

In [26] authors examined local features in the task of detecting manipulated images of human faces and proposed a convolutional architecture and training scheme that target finding such features.

In the [27] authors investigated using deep transfer learning for swapped face detection.

Nhu et al. [28] proposed VGG architecture in application to forensics face detection.

Hsu et al. [29] used a DenseNet-like network for feature extraction and proposed a pairwise learning strategy to enable fake feature learning.

Fernando et al. [30] proposed a Hierarchical Memory Network (HMM) architecture, which can detect faked faces by utilizing knowledge stored in neural memories as well as visual cues to reason about the perceived face and anticipate its future semantic embedding.

III. Face swap detection method

A. Preprocessing

For extraction and classification facial features we use a convolutional neural network. In the first step, every image sends to the preprocessing block. Image preprocessing includes face detection and face alignment.

For face detection we use the method described in [31, 32]. For face alignment we use 5 key points: the outer corners of the eyes, the tip of the nose, and the outer corners of the lips.

An alignment is a necessary step, because the convolutional neural network (which we are going to use for feature extraction) is susceptible to the affine transformations, e.g. rotation, flipping, scaling.

Fig. 2 - Face preprocessing sample: facial point detection result (a); face preprocessing result (b)

Fig. 2 shows the facial point detection result (a) and normalization result (b). The face normalization is conducted in a series of steps. At first, we rotate the face in a horizontal plane by using the outer eyes corners coordinates. It compensates for the tilt of the head to the right and to the left, the rotating about the axis through the nose and back of the head.

In the next step we crop rotated image and scale it. For getting cropping and scaling parameters we use a line between the middle of the eyes and the middle of the mouth, which is invariant to the neck rotation. The cropping area has been chosen to impose the line mentioned above and the horizontal axis center.

In the end, face images were downscaled to 256x256 pixels. Each pixel value in RGB mode is normalized to the range [0,1].

To generate a larger training set we augmented data by random cropping, color, and quality transformation. After random cropping, the image size is 224x224 pixels. Quality transformations include random Gaussian noise, random blur, and random jpeg compression. Color transformations include random contrast, brightness, and gamma correction.

B. Classifier

As a classifier, we use a MobileNet-like convolutional neural network pre-trained on an ImageNet [33, 34]. We transfer it to our task by replacing the final fully-connected layer with two outputs. The other layers are initialized with

the ImageNet weights. We train our networks by PyTorch with mini-batch size 64.

We trained our model using the AdamW stochastic optimization method [35] and softmax cross-entropy loss function.

AdamW was proposed by Lochistov & Hutter and it is improving the Adam regularization [36].

The main idea of the AdamW is decoupling weight decay $d \cdot w_t$ from the gradient and using it directly in the weight update:

$$w_{t+1} = w_t - \eta_t \cdot \left(\frac{m_t}{\sqrt{v_t}+\epsilon} + d \cdot w_t\right) \quad (1)$$

where η_t is a schedule learning rate multiplier, d is weight decay coefficient, m_t- the first-moment vector, v_t- the second-moment vector, which calculated as $m_{t+1} = \beta_1 m_t + (1-\beta_1)g_t$, $v_{t+1} = \beta_2 v_t + (1-\beta_2){g_t}^2$; g_t denotes gradients, β_1, β_2- momentum factors, ε is used to prevent division by 0.

Adam by contrast updates weights using the following formula:

$$w_{t+1} = w_t - \eta_t \cdot \frac{m_t}{\sqrt{v_t}+\epsilon} \quad (2)$$

where η_t is a schedule LR multiplier, m_t, v_t- moment vectors, ε is to prevent division by 0.

As a schedule learning rate multiplier, we use a cosine learning rate annealing (2016) [37]. Cosine Learning Rate Annealing means that within the i-th run, we decay the learning rate for each batch as follows:

$$\eta_t = \eta_{min}^{(i)} + 0.5 \cdot \left(\eta_{max}^{(i)} - \eta_{min}^{(i)}\right) \cdot \left(1 + \cos\left(\frac{\pi \cdot T_{cur}}{T_i}\right)\right), \quad (3)$$

where $\eta_{min}^{(i)}$ and $\eta_{max}^{(i)}$ are ranges for the learning rate, T_{cur} accounts for how many epochs have been performed since the last restart. T_{cur} is updated at each batch iteration t and it is not necessarily an integer value.

The initial learning rate is set to 0.05 and the total amount of epochs is 300. For training we use the FaceForencisc++ dataset [24]. The train set contains 25 278 images, validation set contains 4672 images. The best-performing model was chosen by validation accuracy.

IV. Experiments

A. Dataset

To evaluate the performance of the proposed face swap detection method we use the FaceForencisc++ dataset [24]. FaceForensics++ dataset is a large-scale dataset of manipulated facial imagery.

The dataset contains manipulations created with four state-of-the-art methods, namely, Face2Face, FaceSwap, DeepFakes, and NeuralTextures. It covers three different versions of data based on compression including an original version ($\boldsymbol{C_0}$), slightly-compressed version ($\boldsymbol{C_{23}}$), and heavy-compressed version ($\boldsymbol{C_{40}}$). In this experiment, we only use FaceSwap subset of FaceForensics++ at $\boldsymbol{C_0}$ (high-quality images) and $\boldsymbol{C_{23}}$ (low-quality images) compression levels. Face swap data contains more than 30k images from 1000 videos. The samples of swapping images from the FaceForencisc++ dataset you can see in Fig. 3

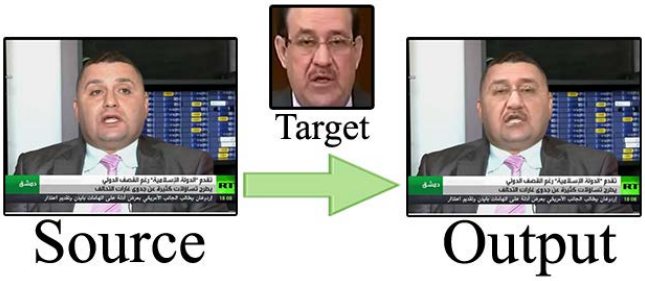

Fig. 3 - Swapping samples from FaceForensics++ dataset.

The convenience of using this dataset is they also propose a state-of-the-art forgery detection method tailored to facial manipulations. It allows not only getting experimental results for our method but also comparing it with other forgery detectors.

B. Result and analysis

As a result of the experiment, the DET (Detection Error Trade-off) curves were made (Fig. 3). The two types of errors were evaluated:

1. False Accept Rate (x-axis): the percentage of classification instances in which swapped faces are incorrectly accepted as original. The formula of FAR is the following:

$$FAR = \frac{number\ of\ misclassified\ real\ face}{total\ number\ of\ real\ face} \quad (N)$$

2. False Reject Rate (y-axis): the percentage of classification instances in which original faces are incorrectly rejected. The formula of FRR is the following:

$$FRR = \frac{number\ of\ misclassified\ fake\ face}{total\ number\ of\ fake\ face} \quad (N)$$

The x-axis is scaled non-linearly by logarithmic transformation to highlight the difference of importance in the critical operating region.

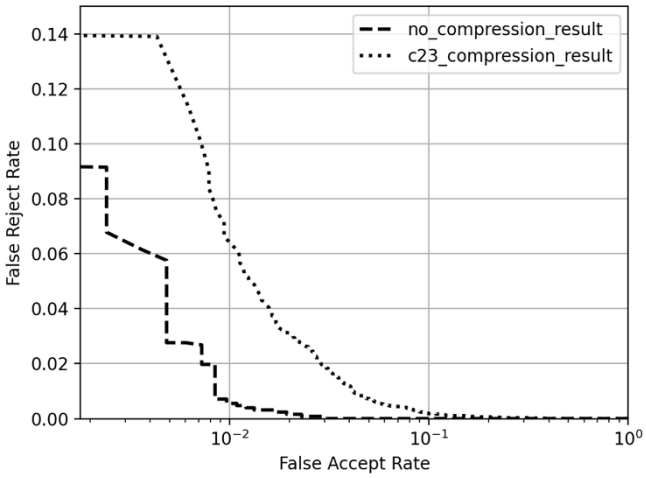

Fig. 4 - DET-curves for face swap detection method on FaceForencics++ dataset with different image compressions

Table 1 shows the accuracies comparison of the different forgery detectors and the proposed method. The accuracy is

used to measure the correct classification performance of the algorithm, and the calculation formula is:

$$Acc = \frac{number\ of\ correct\ classifications}{total\ number\ of\ face} \quad (N)$$

Table 1: Detection accuracies when evaluated on specific compression levels (the bold values are the best performance.

Method	C_0 compression	C_{23} compression
MesoNet [25]	98.15	81.24
Xception [24]	98.39	96.79
Khodabakhsh [23]	**99.11**	-
Tarasiou [26]	-	**98.32**
Wang [38]	-	98.3
Proposed method	**99.28**	**98.31**

According to Table N, our proposed approach shows high level of accuracy for face swap technique detection. We can see a slight decline in this metric when using low-quality images (C_{23} compression). However, our approach outperforms other methods on both levels.

V. Conclusion

In this study, we proposed a method for increasing the security of face authentication systems by detecting swapped faces on the image. The proposed method is comparable with the best results shown when applying the other techniques for detecting swapped faces on the image on the same test dataset. The proposed method is based on a MobileNet-like neural network for getting face features and classifying them for getting the probability of digital tampering using the Face Swap technique. In the study, train technique, hyperparameters, and data used for training and validation were described.

Computational experiments have shown that the proposed method can work with faces obtained with low quality and heavy lighting conditions. It has shown that the accuracy reaches values over 98% for low-quality images and over 99% for high-quality images. The experiments were conducted on one of the biggest face manipulation datasets named FaceForensics++.

Further development of the method for solving the face swap detection attack can be by using recurrent neural networks for additional analysis of connections between several frames.

References

[1] S. Salmani and D. Yadav, "Deepfake: A survey on facial forgery technique using generative adversarial network," 2019 International Conference on Intelligent Computing and Control Systems (ICCS). IEEE, 2019, doi: 10.1109/ICCS45141.2019.9065881.

[2] Vooruit, Officiële Twitter-account van Vooruit, socialistische beweging “Trump heeft een boodschap voor alle Belgen” Twitter, 20 May 2018, Available: https://twitter.com/vooruit_nu/status/998089909369016325.

[3] South China Morning Post, (2021, March 31), “Chinese government-run facial recognition system hacked by tax fraudsters: report,” South China Morning Post, Available: https://www.scmp.com/tech/tech-trends/article/3127645/chinese-government-run-facial-recognition-system-hacked-tax (16 June 2021).

[4] B. Zeno, I. Kalinovskiy, Y. Matveev, and B. Alkhatib, "CtrlFaceNet: Framework for geometric-driven face image synthesis," Pattern Recognition Letters, vol. 138, pp 527-533, 2020, doi: 10.1016/j.patrec.2020.08.026.

[5] J. Zhao, L. Xiong, J. Karlekar, J. Li, F. Zhao, Z. Wang, P. S. Pranata, P. S. Shen, S. Yan, and J. Feng, “Dual-Agent GANs for Photorealistic and Identity Preserving Profile Face Synthesis,” In NIPS, vol. 2, p. 3, Jan. 2017.

[6] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” International Conference on Learning Representations (ICLR), 2018.

[7] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, “Face2Face: Real-Time Face Capture and Reenactment of RGB Videos,” In Conference on Computer Vision and Pattern Recognition (CVPR), 2387-2395, doi: 10.1109/CVPR.2016.262.

[8] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection,” Information Fusion, vol. 64, 2020, pp. 131–148, doi: 10.1016/j.inffus.2020.06.014.

[9] I. Korshunova, W. Shi, J. Dambre, and L. Theis, “Fast face-swap using convolutional neural networks,” In 2017 IEEE International Conference on Computer Vision (ICCV), Oct 2017, pp. 3697–3705, doi: 10.1109/ICCV.2017.397.

[10] Y. Mirsky and W. Lee “The Creation and Detection of Deepfakes: A Survey,” ACM Computing Surveys, vol. 54, iss. 1, April 2021, pp 1–41, doi: 10.1145/3425780.

[11] Y. Li and S. Lyu, “Exposing DeepFake Videos By Detecting Face Warping Artifacts,” In IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019.

[12] A. Agarwal, R. Singh, M. Vatsa, and A. Noore. “SWAPPED! Digital face presentation attack detection via weighted local magnitude pattern,” In 2017 IEEE International Joint Conference on Biometrics (IJCB), IEEE, 2017, pp. 659–665.

[13] R. Durall, M. Keuper, F-J Pfreundt, and J. Keuper, “Unmasking DeepFakes with simple Features,” arXiv preprint arXiv:1911.00686 (2019).

[14] H. Mo, B. Chen, and W. Luo, “Fake faces identification via convolutional neural network,” In Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security, ACM, 2018, pp. 43–47, doi: 10.1145/3206004.3206009.

[15] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, “Face x-ray for more general face forgery detection,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5001–5010, 2020, doi: 10.1109/CVPR42600.2020.00505.

[16] X. Li, K. Yu, S. Ji, Y. Wang, C. Wu, and H. Xue, “Fighting Against Deepfake: Patch&Pair Convolutional Neural Networks (PPCNN),” In Companion Proceedings of the Web Conference 2020, 2020, pp. 88–89, doi: 10.1145/3366424.3382711.

[17] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, “DeepFake Detection Based on the Discrepancy Between the Face and its Context,” arXiv preprint arXiv:2008.12262 (2020).

[18] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, “Protecting World Leaders Against Deep Fakes,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 38–45.

[19] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, “Emotions Don’t Lie: A Deepfake Detection Method using Audio-Visual Affective Cues,” arXiv preprint arXiv:2003.06711 (2020).

[20] D. Guera and E. J. Delp, “Deepfake video detection using recurrent neural networks,” In IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 2018, pp. 1–6, doi: 10.1109/AVSS.2018.8639163

[21] R. Wang, L. Ma, F. Juefei-Xu, X. Xie, J. Wang, and Y. Liu, “Fakespotter: A simple baseline for spoofing ai-synthesized fake faces,” arXiv preprint arXiv:1909.06122 (2019).

[22] S. Fernandes, S. Raj, R. Ewetz, J. S. Pannu, S. K. Jha, E. Ortiz, I. Vintila, and M. Salter, “Detecting Deepfake Videos Using Attribution-Based Confidence Metric,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 308–309, doi: 10.1109/CVPRW50498.2020.00162.

[23] A. Khodabakhsh and C. Busch “A generalizable Feepfake detector based on neural conditional distribution modeling,” 2020 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE, 2020.

[24] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to detect manipulated facial images,” In International Conference on Computer Vision (ICCV), 2019, doi: 10.1109/ICCV.2019.00009.

[25] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “Mesonet: a compact facial video forgery detection network,” arXiv preprint arXiv:1809.00888 (2018).
[26] M. Tarasiou and S. Zafeiriou, “Extracting deep local features to detect manipulated images of human faces,” In 2020 IEEE International Conference on Image Processing (ICIP), 2020, pp. 1821-1825, doi: 10.1109/ICIP40778.2020.9190714.
[27] X. Ding, Z. Raziei, E. C. Larson, E. V. Olinick, P. Krueger, and M. Hahsler, “Swapped Face Detection using Deep Learning and Subjective Assessment,” arXiv preprint arXiv:1909.04217 (2019).
[28] N.T. Do, I. S. Na, and S. H. Kim, “Forensics Face Detection From GANs Using Convolutional Neural Network,” In ISITC’2018, 2018, pp. 376-379.
[29] C.C. Hsu, Y.X. Zhuang, and C.Y. Lee, “Deep fake image detection based on pairwise learning,” Applied Sciences, vol. 10, no. 1:370, 2020, doi: 10.3390/app10010370.
[30] T. Fernando, C. Fookes, S. Denman, and S. Sridharan, “Exploiting Human Social Cognition for the Detection of Fake and Fraudulent Faces via Memory Networks,” arXiv preprint arXiv:1911.07844 (2019).
[31] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, “S^3FD: Single shot scale-invariant face detector,” In ICCV, 2017, doi: 10.1109/ICCV.2017.30.
[32] X. Tang, D. K. Du, Z. He, and J. Liu, “Pyramidbox: A context-assisted single shot face detector,” In ECCV, 2018, doi: 10.1007/978-3-030-01240-3_49.
[33] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, F. F. Li, “ImageNet: A Large-Scale Hierarchical Image Database”, In CVPR09, 2009, doi: 10.1109/CVPR.2009.5206848.
[34] M. Sandler, A. Howard, M. Zhu A. Zhmoginov, and L. Chen, “MobileNetV2: inverted residuals and linear bottlenecks,” In Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.
[35] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” 7th International Conference on Learning Representations, ICLR, 2019.
[36] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” In ICLR, 2015.
[37] I. Loshchilov and F. Hutter, “SGDR: Stochastic gradient descent with warm restarts,” In International Conference on Learning Representations, 2017.
[38] G. Wang, J. Zhou, and Y. Wu Exposing Deep-faked Videos by Anomalous Co-motion Pattern Detection,” arXiv preprint arXiv:2008.04848 (2020)