Theme Artice: Adversarial Examples in CE: Deceptive Threats to AI Systems

Deep Learning-Based Face Forgery Detection for Facial Payment Systems

Siyou Guo
Shandong University of Technology

Qilei Li
Queen Mary University of London

Mingliang Gao, Guisheng Zhang, Jinfeng Pan
Shandong University of Technology

Gwanggil Jeon
Shandong University of Technology
Incheon National University

***Abstract*—The development of Generative Adversarial Networks (GANs) has revolutionized image generation and editing. However, the capacity to create realistic images presents serious security concerns, particularly in the context of face-based payment systems. Deepfakes leverages GANs to generate manipulated videos or images, which may present opportunities for identity theft and fraudulent transactions. For instance, perpetrators employ Deepfakes technology to forge identifying information about victims, such as transplanting their faces into fake videos or images to make it appear like they are performing activities they have never done before. To address this growing concern, this study proposes a deep learning-based detection method utilizing an improved convolutional neural network (CNN) model. The proposed model comprises two key modules, namely the Multi-scale Attention (MA) module and the Halo Attention (HA) module. Specifically, MA is designed to recognize faces and other details in the forged image. HA is built to focus on localized regions of the image. Experimental results show that the proposed model scores 97.12 and 99.32 on FF++ (HQ) dataset and 91.26 and 95.43 on FF++ (LQ) dataset in terms of ACC and AUC, respectively. The remarkable accuracy and performance make it a dependable solution for safeguarding face payment systems.**

FACIAL PAYMENT SYSTEMS have gained widespread adoption across various domains like retail, finance, and public transportation. This innovative technology can enhance user experience and transaction efficiency. For instance, facial payments enable contactless transactions and minimize the risk of virus and bacteria transmission. Furthermore, eliminating physical cards through facial payments

Digital Object Identifier 10.1109/MCE.YYYY.Doi Number
Date of publication DD MM YYYY; date of current version DD MM YYYY

Published by the IEEE Consumer Technology Society

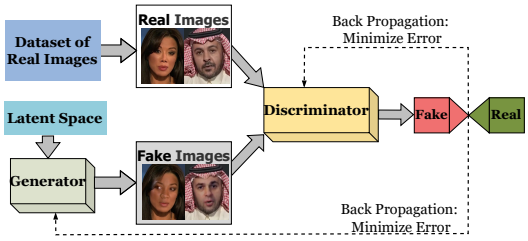

Figure 1. GANs network architecture.

streamlines the payment process and enhances user convenience. However, the burgeoning field of deepfakes technology threatens the facial payment security [1]. It results in unauthorized transactions and the exposure of sensitive user information. Although some financial systems employ security measures, such as detecting vital signs like blinking or head movement, the continuous evolution of deep forgery techniques may breach these security measures.

Deepfakes is a technology that utilizes artificial intelligence techniques, *e.g.*, GANs [2] to produce highly realistic fake videos or images. The working principle of GANs is illustrated in Figure 1. It comprises two modules, *i.e.*, a generator that synthesizes new data from a target image or video, and a discriminator that attempts to distinguish the generated content from real data. Through an iterative refinement process, the generator progressively improves its forgeries to deceive the discriminator, ultimately producing highly convincing deepfakes. This poses a significant challenge to facial recognition systems, as it becomes increasingly difficult to differentiate between a real person and a deepfake.

To detect face forgery in consumer electronics, this work proposes a deep learning-based approach for forgery face detection. The overall framework of the proposed MAHA-Net is shown in Figure 2. Two modules are introduced in this network *i.e.*, the Multi-scale Attention (MA) module is designed to recognize faces and other details in the forged image. The Halo Attention (HA) module is built to focus on localized regions of the image, which helps to recognize subtle changes in expressions and movements of faces in forged images. The main contribution of this paper is as follows:

- To improve the detection accuracy of the deepfake model, a deep leaning-based MAHA-Net is proposed by incorporating a multi-scale Attention (MA) module and a Halo Attention (HA).
- A multi-scale attention is introduced to capture forgery traces at various scales. Meanwhile, a Halo self-attention model is adopted to capture useful relationships between nearby pixels.
- Comparative experiments conducted on public datasets verify that the proposed model outperforms the state-of-the-art methods in detection accuracy.

The remainder of this article is structured as follows. The "Related Work" Section reviews existing methods used for deepfake detection. The "The Proposed Model" Section delves into the details of the proposed deep learning-based method for deepfake detection. The "Experimental Results and Analysis" Section describes the experimental setup, the datasets used, and the evaluation of the performance of the proposed method. Finally, the main conclusions of this work with some possible future trends are given in the "Conclusion and Future Work" Section.

RELATED WORK

Deep forgery detection is to identify and verify the authenticity of deeply forged images or videos, and researchers have given it considerable attention.

Early deepfake detection methods are mainly based on facial cues, *e.g.,* head movements and facial expressions. Jung *et al.* [3] detected deepfakes by analyzing significant changes in the pattern of eye blinking. Ciftci *et al.* [4] leveraged biological signals as implicit authenticity descriptors to detect synthetic content in portrait videos.

Recently, deep learning has dominated the field of deepfake detection [5], [6]. Zhou *et al.* [7] proposed a two-stream neural network for deepfake detection. Specifically, one branch is to analyze the visual appearance and the other focuses on local noise patterns. Rössler *et al.* [8] boosted the performance of deepfake detection by retraining an XceptionNet on manipulated face datasets. Li *et al.* [9] proposed a frequency-aware discriminative feature learning framework for face forgery detection. Nguyen *et al.* [10] proposed a capsule network to improve the deepfake detection accuracy. Cao *et al.* [11] proposed a Reconstruction-classification Learning (RECCE) framework, which learns differences in real face images through reconstruction.

THE PROPOSED MODEL

The proposed MAHA-Net framework is depicted in Figure 2. Compared with baseline RECCE [11],

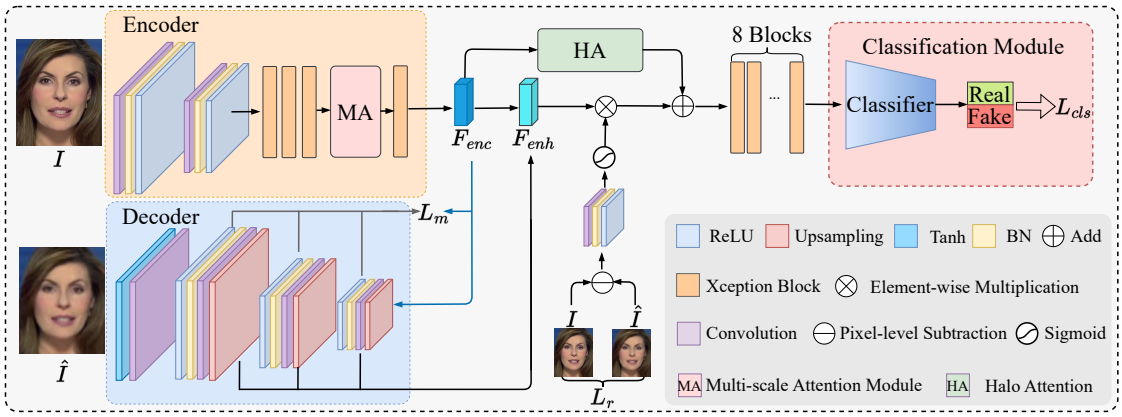

Figure 2. Overall framework of the proposed MAHA-Net.

the Multi-scale Attention (MA) module and the Halo Attention (HA) module are employed in this network. The reconstruction network consists of an encoder and a decoder. Initialization of the encoder involves utilizing a pre-trained Xception model.

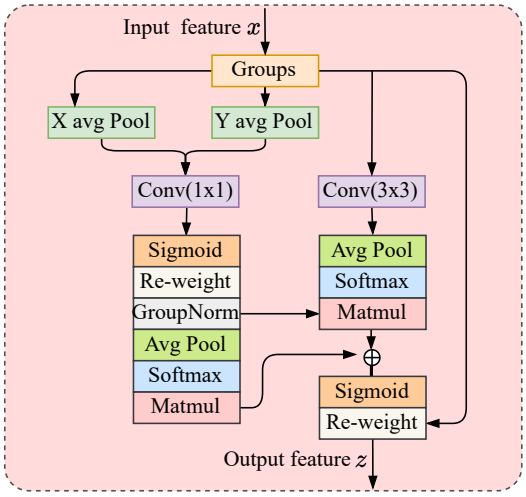

Figure 3. Schematic diagram of the Multi-scale Attention (MA) module.

Within the encoder, the MA module is positioned between the two block layers of the Xception. The input image I is initially processed by the encoder to extract feature maps, denoted by F_{enc}. Following this, F_{enc} is directed through two separate pathways. In one pathway, F_{enc} undergoes a multi-scale graph reasoning operation within the decoder to obtain an enhanced feature representation, denoted by F_{enh}. Simultaneously, the reconstruction differences are utilized to guide the subsequent classification process. The other pathway involves applying the HA module to the feature layer F_{enc}. Finally, the feature maps obtained from both pathways are summed up and used as the basis for the Classification Module, which discriminates between real and fake content.

The output of the image after the reconstruction network is $\hat{I}$. The addition of noise enlarges the coding region of the image, thus masking out the distorted blank coding points. The reconfiguration network equation is as follows:

$$\hat{I} = f_{rec}(\tilde{I}), \tag{1}$$

where the variable $\tilde{I}$ denotes the output obtained by introducing white noise during the training period. $f_{rec}(\cdot)$ denotes the reconstruction network process.

Multi-scale Attention Module

The multi-scale attention (MA) module integrates both channel attention and spatial attention. Therefore, it enables the simultaneous learning of inter-channel and spatial feature dependencies. The architecture of the MA is shown in Figure 3. “Groups” indicates the divided groups, “X Avg Pool” means the 1D horizontal global pooling, and “Y Avg Pool” represents the 1D vertical global pooling, respectively. “Matmul” block is a matrix multiplication operation used to compute interactions between different features.

To acquire various semantic information, the MA module partitions the input image feature map x into g groups along the channel dimension. To capture features at various spatial scales, the MA module employs a three-parallel path structure. Two pathways utilize 1×1 convolutions to focus on fine-grained details within the local area of each pixel. The third pathway employs a 3×3 convolution to capture broader spatial relationships between pixels across a wider area.

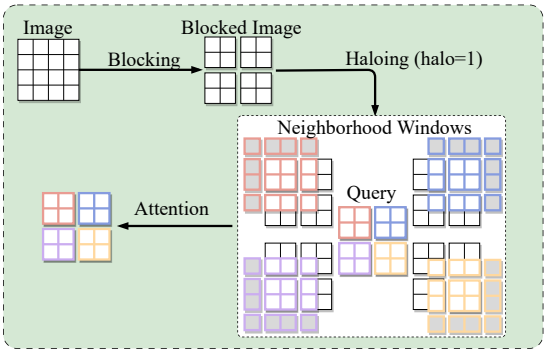

Figure 4. Schematic diagram of the Halo Attention (HA) module.

Halo Attention Module

The halo attention [12] concentrates on a smaller localized window around each pixel. This module enables the proposed framework to capture useful relationships between nearby pixels. The architecture of the halo attention is shown in Figure 4.

Specifically, the input feature map is partitioned into four equally sized blocks. Each block is subjected to a padding operation, *i.e.*, n layers of halos are added around each block. This augmentation expands the perceptual field of each block, and it results in a larger perceptual field for each. Subsequently, each block is individually sampled. Ultimately, the feature map is output and combined with the initial input feature map. This seamless integration is facilitated by the utilization of the residual jump connection, which effectively amalgamates both local and global details.

Loss Function

The overall loss function is composed of three key components, namely reconstruction loss ($\mathcal{L}_{\mathrm{r}}$), classification loss ($\mathcal{L}_{\mathrm{cls}}$), and metric learning loss ($\mathcal{L}_{\mathrm{m}}$). It is formulated as:

$$\mathcal{L} = w_1\mathcal{L}_{\mathrm{m}} + w_2\mathcal{L}_{\mathrm{r}} + \mathcal{L}_{\mathrm{cls}}, \tag{2}$$

where the metric learning loss $\mathcal{L}_{\mathrm{m}}$ minimizes the distance between real data and maximizes the separation between real and synthetic representations in the embedding space. Following the setup described in RECCE [11], the values of w_1 and w_2 are set to 0.1.

EXPERIMENTAL RESULTS AND ANALYSIS

This section details the dataset used, the experimental setup employed, and the analysis of the obtained results. Comparative experiments were conducted on FaceForensics++ (FF++) [8] dataset. FF++ dataset is a forensics dataset comprising 1000 original video sequences manipulated with four automated face manipulation methods. Each video is processed using one of four state-of-the-art face synthesis methods, namely Deepfakes, Face2Face, FaceSwap, and NeuralTextures. To evaluate the performance of the proposed method, we used two metrics, namely Accuracy (ACC) and Area Under the Curve (AUC). The ACC is the ratio of correctly predicted samples to the total number of samples. It provides a straightforward measurement of the model's ability to classify both real and fake images. However, ACC can be misleading with class imbalance. The AUC is a more robust metric that is not affected by class imbalance. Higher ACC and AUC indicate better discrimination ability of the model. We trained the model with 32 samples per batch for 40 iterations. The learning rate is dynamically adjusted during training using a step-learning rate scheduler. We implemented the framework using PyTorch [13] and trained it on two 3090 Ti GPUs in parallel.

Table 1. Comparison with the SOTA methods. The best results are highlighted in bold.

Methods	FF++ (HQ)		FF++ (LQ)	
	ACC	AUC	ACC	AUC
SPSL [14]	91.50	95.32	81.57	82.82
RFM [15]	95.69	98.79	87.06	89.83
Multi-task [16]	85.65	85.43	81.30	75.59
Xception [17]	95.73	96.30	86.86	89.30
Add-Net [18]	96.78	97.74	87.50	91.01
Two-branch [19]	96.43	98.70	86.34	86.59
Freq-SCL [9]	96.69	99.28	89.00	92.39
RECCE [11]	97.06	**99.32**	91.03	95.02
MAHA-Net (Ours)	**97.12**	**99.32**	**91.26**	**95.43**

As shown in Table 1, the proposed method performs best on FF++ datasets. Compared to the baseline RECCE [11], the proposed method achieves the same excellent results on FF++ (HQ) dataset. On FF++ (LQ) dataset, the proposed model improves the ACC and AUC to 91.26% and 95.43%, respectively. These results indicate that the proposed network has an advantage in accurately detecting face forgery images compared with the SOTA competitors. The performance improvement can be attributed to the proposed MA and HA modules. Specifically, the MA module captures and integrates features across multiple scales, which enhances the ability of the network to detect

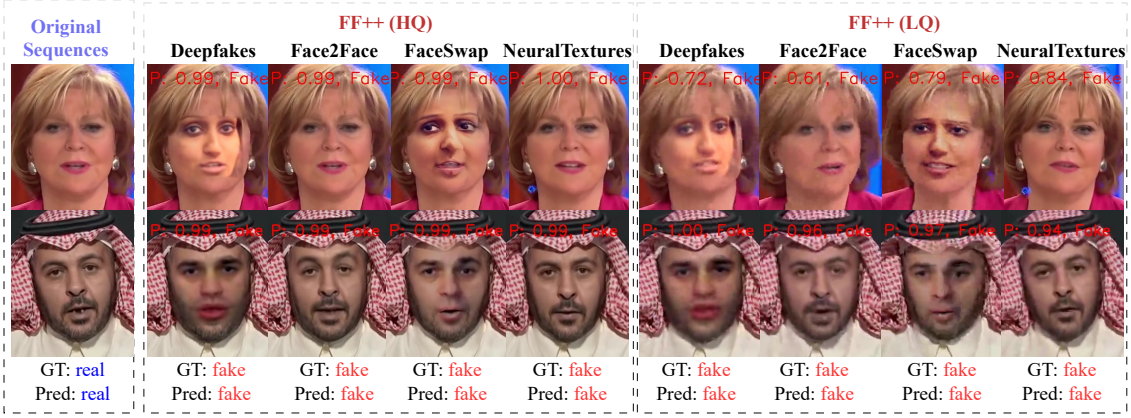

Figure 5. The inference results of deepfake detection by the proposed model.

subtle and varied patterns in deepfake images. The HA module focuses on the most relevant regions. It enables the model to distinguish between authentic and manipulated content and thus boost performance, especially in low-quality datasets. Low-quality data is susceptible to adversarial attacks [20]. This is because the noise and artifacts present in low-quality data can provide a hiding place for adversarial perturbations. Compared to the SOTA methods, the superior performance of the proposed framework on low-quality datasets verifies the robustness of the proposed model.

The objective result of deepfake detection by the proposed model is shown in Figure 5. It shows that the proposed MAHA-Net can accurately predict the authenticity of human faces.

CONCLUSION AND FUTURE WORK

Deep face forgery poses a significant security risk to face payment systems in the consumer electronics domain. To mitigate these risks, a deep forgery detection framework named MAHA-Net is presented in this work. MAHA-Net integrates a multi-scale attention module and a halo attention module. The multi-scale attention module captures image features at different scales, and the halo attention module focuses on localized regions of the image. The two modules can extract the forged features in deep forged images. Extensive experiments have demonstrated that MAHA-Net outperforms current state-of-the-art methods in accuracy on the widely used FF++ dataset.

The results indicate that the detection accuracy of the MAHA-Net and existing detection methods was low on the low-quality FF++ datasets. In the future, we will aim to improve the accuracy of low-pixel forged faces. Additionally, we plan to investigate the integration of techniques like transfer learning to further enhance the robustness of MAHA-Net against new and unseen deepfakes.

ACKNOWLEDGMENTS

This work was supported partly by the National Natural Science Foundation of China (No.61801272)

REFERENCES

1. S. Saeedi, A. C. Fong, S. P. Mohanty, A. K. Gupta, and S. Carr, “Consumer artificial intelligence mishaps and mitigation strategies,” *IEEE Consumer Electronics Magazine*, vol. 11, no. 3, pp. 13–24, 2021.
2. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
3. T. Jung, S. Kim, and K. Kim, “Deepvision: Deepfakes detection using human eye blinking pattern,” *IEEE Access*, vol. 8, pp. 83 144–83 154, 2020.
4. U. A. Ciftci, I. Demir, and L. Yin, “Fakecatcher: Detection of synthetic portrait videos using biological signals,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 01, pp. 1–1, 2020.
5. F. Ding, G. Zhu, M. Alazab, X. Li, and K. Yu, “Deep-learning-empowered digital forensics for edge consumer electronics in 5g hetnets,” *IEEE consumer electronics magazine*, vol. 11, no. 2, pp. 42–50, 2020.
6. G. Zhang, M. Gao, Q. Li, W. Zhai, G. Zou, and G. Jeon, “Disrupting deepfakes via union-saliency adversarial attack,” *IEEE Transactions on Consumer Electronics*, vol. 70, no. 1, pp. 2018–2026, 2024.
7. P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, “Two-stream neural networks for tampered face detection,” in *2017 IEEE conference on computer*

vision and pattern recognition workshops (CVPRW). IEEE, 2017, pp. 1831–1839.

8. A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.
9. J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6458–6467.
10. H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2019, pp. 2307–2311.
11. J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4113–4122.
12. A. Vaswani, P. Ramachandran, A. Srinivas, N. Parmar, B. Hechtman, and J. Shlens, "Scaling local self-attention for parameter efficient visual backbones," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 894–12 904.
13. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS-W*, 2017.
14. H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 772–781.
15. C. Wang and W. Deng, "Representative forgery mining for fake face detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 923–14 932.
16. H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," in *2019 IEEE 10th international conference on biometrics theory, applications and systems (BTAS)*. IEEE, 2019, pp. 1–8.
17. F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.
18. B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, "Wilddeepfake: A challenging real-world dataset for deepfake detection," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 2382–2390.
19. I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*. Springer, 2020, pp. 667–684.
20. Y. Lu and T. Ebrahimi, "Assessment framework for deepfake detection in real-world situations," *EURASIP Journal on Image and Video Processing*, vol. 2024, no. 1, p. 6, 2024.

Siyou Guo is currently working toward an M.S. degree with the School of Shandong University of Technology, Zibo, China. Contact him at 23504030565@stumail.sdut.edu.cn.

Qilei Li is working toward a PhD at the Queen Mary University of London, London, E1 4NS, United Kingdom. Qilei Li and Siyou Guo contributed equally to this work. Contact him at qilei@ieee.org.

Mingliang Gao is an associate professor and vice dean at the Shandong University of Technology. He is the first corresponding author of this article. Contact him at mlgao@sdut.edu.cn.

Guisheng Zhang is working toward an M.S. degree with the School of Shandong University of Technology, Zibo, China. Contact him at 22504030001@stumail.sdut.edu.cn.

Jinfeng Pan is an associate professor at the School of Electrical and Electronic Engineering, Shandong University of Technology, Zibo, China. Contact her at pjfbysj@163.com.

Gwanggil Jeon is a professor at Shandong University of Technology, Zibo, China, and Incheon National University, Incheon, Korea. He is the second corresponding author of this article. Contact him at ggjeon@gmail.com.