Theme Article: Cognitive and Semantic Computing for Enhanced User Experience

Towards Sequential Deepfake Detection Using Deep Learning for Privacy Protection

Guisheng Zhang
Shandong University of Technology

Mingliang Gao
Shandong University of Technology

Qilei Li
Queen Mary University of London

Gwanggil Jeon
Shandong University of Technology
Incheon National University

***Abstract*—With the development of deep learning, deep generative models have generated hyper-realistic facial images that are virtually indistinguishable from authentic images. Recently, the misuse of deepfake technology in electronic consumption is becoming increasingly prevalent. This poses a significant threat to consumer privacy and property security. In reaction to this issue, an array of methods for deepfake detection has been proposed to evaluate the authenticity of images. Sequential deepfake detection is an extension of the deepfake detection approach. It aims to detect various facial manipulation operations and accurately identify the sequence of facial manipulations. To enhance the accuracy of sequential deepfake detection and protect consumer privacy, we propose a deep learning-based detection method for sequential deepfake detection. It is designed to extract fine-grained features for detecting facial manipulation sequences. Compared to the state-of-the-art methods, the proposed method improves the Fixed-Acc and Adaptive- Acc metrics by 1.43% and 3.89%, respectively.**

THE RAPID ADVANCEMENT of cognitive and semantic computing has engendered a wave of novel intelligent applications in the digital era. Specifically, the application of certain semantic-enabled consumer electronics in digital media has brought significant convenience to consumers. However, the rampant misuse of deepfakes in the electronic consumption landscape poses a significant threat to individual privacy and financial security [1], [2]. For example, some individuals acquire consumers' facial images by using semantic consumer electronics. The semantic consumer electronic refers to consumer electronic devices that are capable of understanding and processing human language and context through the use of semantic

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

technologies. Then, they employ a deepfake model to create fake images. Subsequently, these forged images are employed in scams during electronic transactions.

Deepfake is a technique for synthesizing images. It generates deceptive and misleading media content by utilizing deep learning and artificial intelligence algorithms. Deepfake generation methods are commonly used to create convincingly false images. The rapid development of these methods has raised concerns about the misuse of false information and privacy infringement. It has become imperative to explore effective technological solutions to mitigate the risks posed by deepfake methods in electronic consumption and to protect consumer privacy and financial security. Therefore, many researchers are dedicated to studying and proposing solutions to address this issue [3], [4], [5]. Deepfake detection is a crucial area of research for addressing these challenges. Deepfake detection is a method that employs algorithms and technical tools to discern and identify deepfake media. Common methods for deepfake detection encompass deep learning-based technologies along with conventional digital forensics and digital signal processing methodologies.

Facial images may undergo a series of forged operations (*e.g.*, changing the hair color, nose shape, and eye size), each with its specific sequence. For example, in Figure 1 (b), the original image is manipulated sequentially to obtain a fake image. The operation sequence of the fake image is “Eyebrow-Nose”. Nevertheless, traditional deepfake detection only identifies the authenticity of images. They lack the ability to detect manipulation sequences.

To address this problem, a Seq-DeepFake Transformer (SeqFakeFormer) [6] was proposed to detect sequences of facial manipulations. This network effectively mitigates the risks associated with deepfake technology in electronic consumption. The SeqFakeFormer employs an improved autoregressive model to identify manipulated regions. However, The effective detection of facial manipulations necessitates the meticulous extraction of features from regions of varying sizes. To solve this issue, we propose a multifaceted attention and spatial-frequency attention network for sequential deepfake detection. The multifaceted attention enables the proposed method to capture global and local features. The spatial-frequency attention enhances the capability of the proposed method to capture finely subtle features within deepfake images. The main contribution is as follows:

- To improve the detection accuracy of the sequential deepfake model, an MASFA-Net is proposed by incorporating a multifaceted attention module and a spatial-frequency attention module.
- A multifaceted attention is employed to capture global and local features. Meanwhile, a spatial-frequency attention module is proposed to capture finely subtle features within fake images.
- Experiments verify that the proposed model outperforms the state-of-the-art methods in accuracy.

The rest of this article is organized as follows. The “Related Work” Section presents related work on deepfake detection, sequential deepfake detection, and attention mechanisms. The “Proposed Sequential Deepfake Detection Method” Section illustrates the proposed method in detail. The “Experimental Results and Analysis” Section analyses the experimental results. Finally, the main conclusions of this work are given in the “Conclusion” Section.

RELATED WORK

Deepfake detection refers to the process of using technological methods to identify forged videos or images generated by deep learning algorithms. Sequential deepfake detection is a recently emerging research area that identifies deepfakes created through various manipulations. Both deepfake detection and sequential deepfake detection share the common objective of differentiating between authentic and manipulated images. However, sequential deepfake detection aims to identify and localize manipulated facial regions. Meanwhile, it detects the sequence of these manipulations. The distinction between deepfake detection and sequential deepfake detection is illustrated in Figure 1. Recently, a Seq-DeepFake Transformer (SeqFakeFormer) [6] was proposed to detect sequences of facial manipulations. The SeqFakeFormer employs an improved autoregressive model to identify manipulated regions. However, the effective detection of facial manipulations necessitates the meticulous extraction of features from regions of varying sizes. Current sequential deepfake detection methods fall short in this capacity. In this work, we propose the MASFA-Net to capture finely subtle features within fake images and improve the accuracy of sequential deepfake detection.

The attention mechanism is designed to enhance the network to focus on different parts of the input. It has been applied across the computer vision domain. Many attention models have been proposed, *e.g.,* Multifaceted Attention [7], Triplet Attention [8], etc. The

This article has been accepted for publication in IEEE Consumer Electronics Magazine. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MCE.2024.3416905

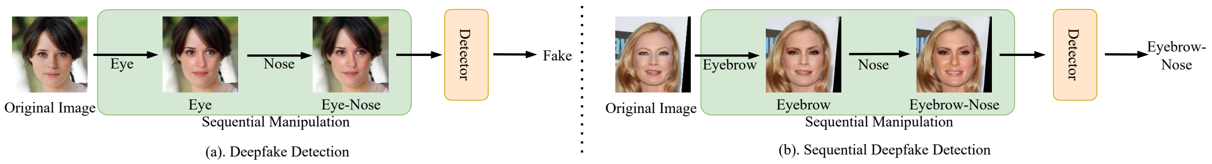

Figure 1. Comparison between (a) deepfake detection and (b) sequential deepfake detection.

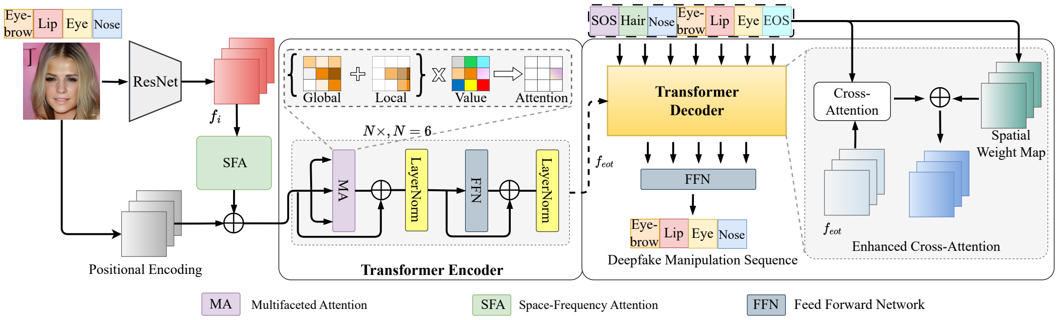

Figure 2. Overall framework of the proposed model.

attention mechanism enables the deepfake detection model to focus on key regions while processing facial images. In this work, we introduce the multifaceted and spatial-frequency attention modules to enhance model accuracy in sequential deepfake detection.

PROPOSED METHOD

In this work, a Multifaceted Attention and Spatial-Frequency Attention Network (MASFA-Net) is employed to improve the accuracy of sequential deepfake detection. The framework of the MASFA-Net is shown in Figure 2. Specifically, an input image is fed into the ResNet50 [9] to extract the feature f_i. Then, a spatial-frequency attention module analyzes the feature f_i to focus on important areas of the face. The spatial-frequency attention module can concentrate on spatial and frequency information concurrently. Consequently, it considers the spatial position of pixels and features at different frequencies during image data processing. Subsequently, a transformer encoder is employed to extract spatial relationships among operational region features. The output of the transformer encoder is defined as f_{eot}. In contrast to the traditional transformer encoder, a multifaceted attention module is introduced to substitute the original self-attention mechanism in MASFA-Net. Then, different forgery operations are translated into distinct tokens. Start of Sequence (SOS) and End of Sequence (EOS) tokens are inserted at the beginning and end of the sequence composed of these forgery operations. Next, a tokenized sequence of operations is obtained. Finally, the feature f_{eot} and a tokenized sequence of operations are fed to the transformer decoder. Sequential manipulation traces are captured by modeling sequential relations based on spatial features. This is achieved through cross-attention modules in the decoder with an autoregressive mechanism. To improve the network's performance with limited labelled manipulation sequences, we follow the SeqFakeFormer method and use a spatially enhanced cross-attention module. This module generates different spatial weight maps for corresponding manipulations to improve cross-attention. The proposed method is trained using a cross-entropy loss.

Spatial and frequency information are important in computer vision. Spatial information excels in processing an image's overall structure and semantic information. Frequency information excels in handling an image's local details and texture information. Therefore, spatial-frequency attention is employed to capture fine-grained facial features. The structure of the spatial-frequency attention is shown in Figure 3.

The spatial-frequency attention module comprises three branches, each incorporating spatial and frequency information extraction modules. The modules for spatial and frequency information extraction are in a parallel state. Given an input tensor with shape $(C \times H \times W)$, each branch aggregates cross-dimensional interaction features between the spatial dimensions H or W and the channel dimension C. This is achieved by straightforwardly rearranging the

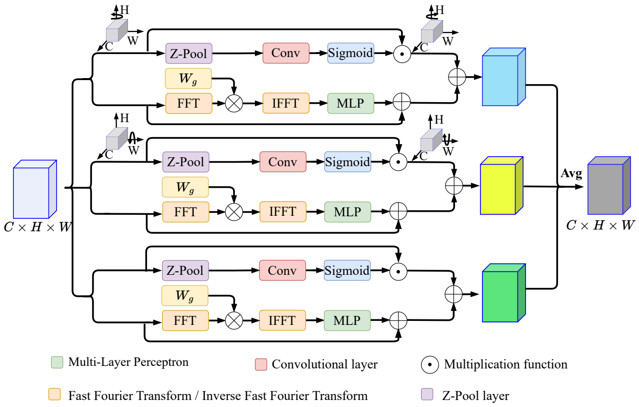

Figure 3. Schematic diagram of the Spatial-frequency Attention module.

input tensors in each branch. Then, a Z-pool layer is applied to pass the tensors, and a convolutional layer with a kernel size of $k \times k$ is applied. The Z-pool layer reduces the zeroth dimension of the tensor to two by concatenating the average-pooled and max-pooled features across their dimension. Subsequently, attention weights are generated through a sigmoid activation layer and applied to the permuted input tensor. Finally, the tensor is transformed back to its original input shape.

Each branch has identical spectral layers for the frequency information extraction part. The spectral layer assesses image frequencies using a spectral gating network comprising a Fast Fourier Transform (FFT) layer, weighted gating with a trainable parameter W_g, and an Inverse FFT layer (IFFT). The spectral layer converts the physical space into spectral space by employing FFT. The trainable weight parameter W_g facilitates precise weight assignment to each frequency component. The W_g ensures the accurate capture of lines and edges within an image. The acquisition of this parameter is accomplished through the implementation of back-propagation techniques. Subsequently, the IFFT is applied to revert the spectral space to the physical space. The Multi-Layer Perceptron (MLP) is used for channel mixing. Finally, the spatial features and frequency features are fused. The output of this module is obtained by averaging the outputs of the three branches.

The self-attention module is pivotal in many deep learning models, especially transformers. However, self-attention still falls short of adequately capturing local information. Therefore, we employ multifaceted attention [7] to extract features of forged traces at different scales. Forged traces refer to subtle anomalies or features in artificially generated media that reveal the content has been tampered with or synthesized. The multifaceted attention comprises global attention and local attention. Global attention is a standard self-attention mechanism employed for extracting global information. It is computed by:

$$Att_g(Q,K,V) = \mathcal{S}(\frac{QK^T}{\sqrt{d}})V, \tag{1}$$

where Q, K, and V refer to Query, Key, and Value. $\mathcal{S}$ is the sigmoid function. For local attention, the goal is to seek a mechanism to learn the most suitable local region for each forged operation. Based on prior knowledge, a rectangular region can be identified by two vertices. We obtain a two-dimensional learnable attention map using these two points. First, given query vectors $Q \in R^{WH \times d}$ and key vectors $K \in R^{WH \times d}$, two predicted coverage probability maps M_1, M_2 can be computed by trough learnable parameter matrices W_1, W_2. The M_1 and M_2 can be obtained as:

$$\begin{aligned} M_1 &= \mathcal{S}((QW_1^Q)(KW_1^K)^T), \\ M_2 &= \mathcal{S}((QW_2^Q)(KW_2^K)^T). \end{aligned} \tag{2}$$

To obtain a two-dimensional learnable attention map, resize M_1 and M_2 into a 2D matrix of size $W \times H$. For each position i along the first axis of M_1 and M_2, there are two corresponding probability maps, M_{1i} and M_{2i}. Subsequently, the learnable region map R is redesigned by applying the cumulative distribution function (CDF) to M_{1i} and M_{2i}. Finally, the learnable region map R is employed for local attention. The local attention can be expressed as:

$$Att_l(Q,K,V) = \mathcal{S}(\frac{QK^T \circ R}{\sqrt{d}})V, \tag{3}$$

where $\circ$ is defined as the Hadamard product [10]. Finally, multifaceted attention is composed by integrating global and local attention mechanisms. The multifaceted attention enables the transformer encoder to adeptly capture intricate spatial relationships within the feature of operational regions.

EXPERIMENTAL ANALYSIS

The Facial Components Dataset is utilized to train the proposed MASFA-Net. This dataset was collected and produced by Shao *et al.* [6]. The dataset encompasses 35,166 manipulated facial images, each meticulously labelled with 28 manipulation sequences of varying lengths. Annotations detailing the sequential operations conducted on facial components accompany each image. The distribution of manipulation sequence

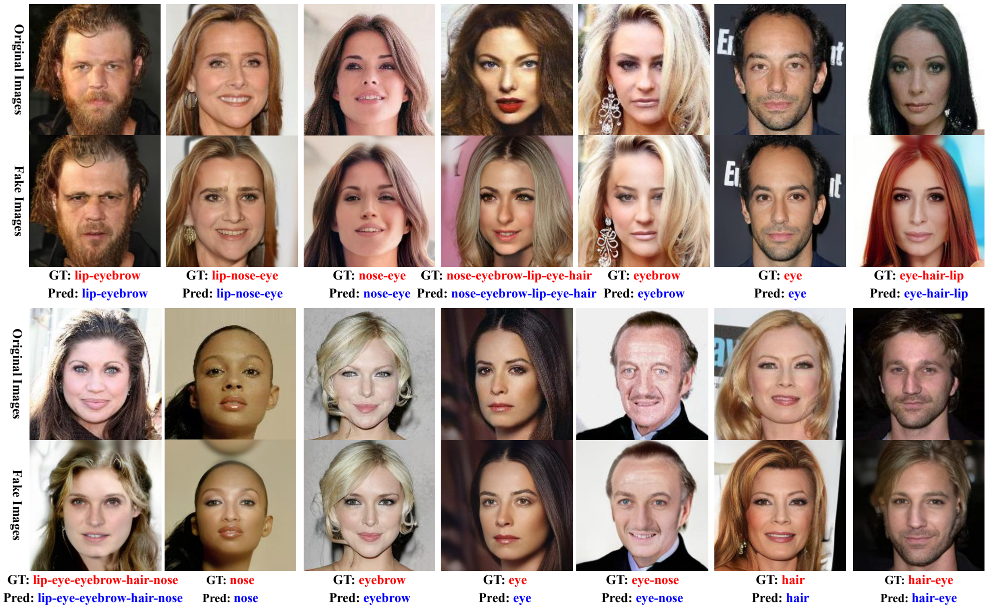

Figure 4. The visualized results of the proposed model. The texts in red and blue signify the ground truth (GT) and the prediction (Pred), respectively.

lengths, spanning from 1 to 5, is delineated as follows: 20.48%, 20.06%, 18.62%, 20.88%, and 19.96%.

To evaluate the performance of the proposed method, we employ two metrics: Fixed Accuracy (Fixed-Acc) and Adaptive Accuracy (Adaptive-Acc). The Fixed-Acc denotes the accuracy when detecting facial manipulation sequences of fixed lengths. Similarly, Adaptive-Acc denotes the accuracy in detecting facial manipulation sequences with adaptive lengths. Higher scores in Adaptive-Acc and Fixed-Acc indicate better network performance.

This work employs a ResNet50 backbone network to perform coarse-grained feature extraction from images. The batch size is set to 64, and the epoch is set to 150. The initial learning rates are set as 1e-3 for the transformer part and 1e-4 for the ResNet50 part. The trainable weight parameter W_g is defined as a matrix with all initial values 1. The framework is implemented in PyTorch [11] with 2 NVIDIA 3090Ti GPUs.

The comparison results between the proposed method and SOAT methods are shown in Table 1. In Table 1, the proposed method performs best on the facial components dataset. Specifically, the proposed method achieves scores of 72.61 and 55.57 in terms of Fixed-Acc and Adaptive-Acc, respectively. Compared to the SeqFakeFormer [6], the proposed method improves the Fixed-Acc and Adaptive-Acc metrics by 1.43% and 3.89%, respectively. However, according to Table 1, the performance of the proposed method is not satisfactory when there is no limit on the number of forgery operations. Figure 4 illustrates the visual outcomes of detecting consecutive deepfake manipulations. The subjective assessments reveal the adept discernment by the proposed neural network of deepfake manipulation sequences characterized by diverse lengths.

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

Methods	facial components	
	Fixed-Acc	Adaptive-Acc
DRN [12]	66.06	45.79
Multi-Cls [6]	69.65	50.57
DETR [13]	69.75	49.84
MA [14]	71.31	52.94
SeqFakeFormer [1] [6]	71.58	53.62
MASFA-Net (ours)	**72.61**	**55.71**

CONCLUSION

In this work, to mitigate the potential risk caused by deepfake in electronic consumption, we proposed multifaceted and spatial-frequency attention networks for sequential deepfake detection. This framework introduces the multifaceted and spatial-frequency attention modules to extract fine-grained features. Compared to existing methods for sequential deepfake

[1] Performance evaluation was conducted with the officially released code and performed on the same platform as ours.

manipulation detection, the proposed method excels in extracting the spatial features of facial manipulation regions at a finer granularity to enhance the detection accuracy of the network. The experiment results demonstrate that the proposed method can enhance the accuracy of sequential deepfake detection and it can safeguard consumer privacy and security. However, when the length of the detected forgery sequence is unfixed, the detection performance of the proposed method is poor. Therefore, our future work will continue to enhance the model's accuracy in detecting sequences of adaptive lengths.

Acknowledgments

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

REFERENCES

1. K. Sharma, A. Malik, I. Batra, A. Sanwar Hosen, M. A. Latif Sarker, and D. S. Han, “Technologies behind the smart grid and internet of things: A system survey.” *Computers, Materials & Continua*, vol. 75, no. 3, 2023.
2. 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.
3. Q. Li, M. Gao, G. Zhang, W. Zhai, J. Chen, and G. Jeon, “Towards multimodal disinformation detection by vision-language knowledge interaction,” *Information Fusion*, vol. 102, p. 102037, 2024.
4. 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, 2023.
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. R. Shao, T. Wu, and Z. Liu, “Detecting and recovering sequential deepfake manipulation,” in *European Conference on Computer Vision*. Springer, 2022, pp. 712–728.
7. H. Lin, Z. Ma, R. Ji, Y. Wang, and X. Hong, “Boosting crowd counting via multifaceted attention,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 628–19 637.
8. D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, “Rotate to attend: Convolutional triplet attention module,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3139–3148.
9. H. M. Nguyen and R. Derakhshani, “Eyebrow recognition for identifying deepfake videos,” in *2020 international conference of the biometrics special interest group (BIOSIG)*. IEEE, 2020, pp. 1–5.
10. R. A. Horn, “The hadamard product,” in *Proc. Symp. Appl. Math*, vol. 40, 1990, pp. 87–169.
11. 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.
12. S.-Y. Wang, O. Wang, A. Owens, R. Zhang, and A. A. Efros, “Detecting photoshopped faces by scripting photoshop,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 072–10 081.
13. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” in *European conference on computer vision*. Springer, 2020, pp. 213–229.
14. H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, “Multi-attentional deepfake detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 2185–2194.

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

Qilei Li is working toward a Ph.D. with the University of Queen Mary University of London, London, E1 4NS, United Kingdom. Qilei Li and Guisheng Zhang contributed equally to this work. Contact him at q.li@qmul.ac.uk.

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.

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.