Deepfake Detection via a Progressive Attention Network

Siyou Guo[1], Mingliang Gao[1], Qilei Li[2], Gwanggil Jeon[3], and David Camacho[4]
[1]Shandong University of Technology, Zibo 255000, China
23504030565@stumail.sdut.edu.cn, mlgao@sdut.edu.cn
[2]Queen Mary University of London, London E1 4NS, United Kingdom
q.li@qmul.ac.uk
[3]Incheon National University, Incheon 22012, South Korea
ggjeon@gmail.com
[4]Universidad Politécnica de Madrid (UPM), Madrid, Spain
david.camacho@upm.es

2024 International Joint Conference on Neural Networks (IJCNN) | 979-8-3503-5931-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IJCNN60899.2024.10650463

***Abstract*—The rapid advancement of deepfake technology has enabled the creation of highly realistic forged face images or videos. While deepfake technology adds entertainment to people's lives, it also poses a potential threat to social security. Deepfake detection is a crucial technology for identifying forged images. However, existing deep learning-based models for deepfake detection often overlook subtle forged traces. To solve this problem, we propose a Progressive Attention Network (PANet). The PANet incorporates two attention modules, namely the Efficient Multi-Scale Attention Module (EMAM) and the Spatial and Channel Attention Module (SCAM), in a progressive manner. The EMAM focuses on crucial facial regions, such as the eyes, nose, and mouth, rather than the entire face. The SCAM facilitates fine-grained feature extraction. Experimental results demonstrate that the proposed method achieves state-of-the-art results on deepfake detection datasets.**

***Index Terms*—Deepfake Detection, Efficient Multi-Scale Attention, Feature Extraction, Information Disorder, Forged Traces**

I. Introduction

The advent of deepfake technology, especially Generative Adversarial Networks (GANs) [1], has made rumor-mongering easier. Deepfakes often manipulate audio and visual elements to depict individuals saying or doing things they never did [2]. Also, deepfakes can be weaponized to spread fabricated narratives and amplify the impact of disinformation campaigns.

Deepfakes typically generate artifacts that may be subtle to humans. However, they can be detected using specified deepfake detection models. Early machine learning face forgery detection methods commonly followed the classical approach of training Convolutional Neural Networks (CNNs) for image classification. These methods used ready-made CNN backbones that directly processed facial images as inputs and classified them as genuine or fake. Nevertheless, these vanilla CNNs lacked an in-depth understanding of forgeries. To address this problem, recent studies have integrated specific forgery patterns, including spectral analysis, noise features, detailed textures, and frequency information. This approach aims to detect traces of forgery in manipulated facial images with greater accuracy and robustness. For instance, Wang *et al.* [3] combined the RGB and frequency domains resulting in improved robustness of the deepfake detection model. Huang *et al.* [4] exploited the spectral artifacts left over from the upsampling operation used in the GAN to detect high-fidelity deep vacation videos. Wang *et al.* [5] introduced a forensic-inspired approach using efficient "Multi-Head Relative-Interaction" to pinpoint deepfakes by analyzing video noise patterns.

The aforementioned methods rely on the forgery patterns associated with specific operational techniques. To solve this problem, some research was devoted to learning representations that can be generalized to unknown forgery patterns. Chen *et al.* [6] dynamically adjusted the model by synthesizing pseudo-training data, thereby enhancing the generality of the deepfake detector against unseen forgeries. Meanwhile, the incorporation of an attention mechanism in deep forgery detection enhances the model's ability [7]. The Reconstruction-classification Learning (RECCE) [8] framework addresses uncertain forgery patterns by learning differences in real face images through reconstruction. However, the RECCE framework exhibits suboptimal performance in the extraction of fine-grained features. To solve this problem, we propose a Progressive Attention Network (PANet) in this paper. The PANet incorporates two attention modules, namely the Efficient Multi-Scale Attention Module (EMAM) [9] and the Spatial and Channel Attention Module (SCAM) [10], in a progressive manner. The main contributions are concluded as follows.

- We propose a deepfake detection method via a Progressive Attention Network. The method uses a progressive attention module to focus on the extracted features from coarse-grained to fine-grained, which improves the accuracy of detecting subtle traces of forgery.
- We introduce an efficient multi-scale attention module. This module is designed for coarse-grained feature learning and focuses on key regions in face features.
- We employ spatial and channel attention modules to learn and emphasize local structures of the face image and extract fine-grained features.

Corresponding authors: Mingliang Gao, David Camacho

Authorized licensed use limited to: Queen Mary University of London. Downloaded on November 04,2024 at 15:34:47 UTC from IEEE Xplore. Restrictions apply.

II. Related Work

A. Deepfake Detection

Deepfake detection is critical for protecting information trustworthiness, privacy, legal social stability, and regulating technological developments. Researchers are working to improve the accuracy and efficiency of deepfake detection. Shao *et al.* [11] recognized successive depth forgery operations by establishing enhanced correspondences between pairs of image sequences. Wodajo *et al.* [12] combined the learning capabilities of CNNs and transformers to detect deepfake video. Heo *et al.* [13] combined vector-cascaded CNN features with patch-based localization to interactively specify artifact regions across all locations. Das *et al.* [14] utilized face marker information to dynamically cut regions in an image. The RECCE [8] framework learns real face image differences by reconstruction to effectively handle uncertain forgery patterns. However, it lacks extracting fine-grained features. To solve this problem, we propose a Progressive Attention Network (PANet).

B. Attention mechanism

The attentional mechanism is a technique that allows a model to learn the relationships between different inputs. There are three widely acknowledged attention mechanisms, namely channel attention (CA), spatial attention (SA), and a combination of CA and SA. The CA is exemplified by the Squeeze-and-excitation (SE) mechanism [15], which explicitly models the relationships between different dimensions and extracts attention according to the channel dimension. Meanwhile, the Convolutional Block Attention Module (CBAM) [10] captures the semantic inter-mapping relationship between spatial and channel dimensions in the feature extraction by incorporating cross-channel and cross-spatial information. The Spatial Group-wise Enhance (SGE) attention [16] enhances the spatial distribution of different semantic sub-features by grouping the channel dimension. The Inverted Residual Mobile Block (iRMB) attention [17] extends the spatial dimension of the input feature map and computes attention weights between features using the multi-head self-attention mechanism.

Considering that the key facial features, *e.g.,* eyes, nose, and mouth are typically the most recognizable elements of the face, they are the most susceptible to manipulation in the context of deepfakes. To effectively deepfake detection, attention mechanisms can be utilized to focus on these critical facial components [8], [18].

III. Proposed method

A. Overview

In this work, we propose a Progressive Attention Network (PANet) by equipping the RECCE framework with an Efficient Multi-Scale Attention Moudle (EMAM) [9]. The EMAM focuses on the important feature representation of images and achieves coarse-grained feature extraction. At the same time, we adopt a Spatial and Channel Attention Module (SCAM) to further refine feature extraction, thereby achieving a coarse-to-fine feature extraction. The structure of the proposed Progressive Attention Network framework is depicted in Fig. 1.

Given an input image $X \in \mathbb{R}^{C\times H\times W}$, it is first fed into the encoder initialized by Xception, which extracts common features. The output of the image after the reconstruction network is $\hat{X}$. The introduction of noise enlarges the coding region of the image, thus masking out the distorted blank coding points [19]. The common feature extraction equation is as follows:

$$\hat{X} = \mathrm{F_{xcep}}(\tilde{X}), \tag{1}$$

where $\tilde{X}$ denotes the result of adding white noise during the training period.

B. Efficient Multi-Scale Attention Module

The Efficient Multi-Scale Attention Module (EMAM) consists of three parts, namely feature grouping, parallel subnetworks, and cross-spatial learning. Feature grouping divides the input feature map into sub-features, parallel subnetworks extract local and global information from the grouped feature map. Cross-spatial learning is employed to aggregate attention-weight descriptors from the parallel subnetworks to capture the pairwise relationships between pixels in the input image. The architecture of the EMAM is shown in Fig. 2. To acquire various semantic information, for the input image feature map $X \in \mathbb{R}^{C\times H\times W}$, the EMAM partitions the feature maps into G groups within the channel dimension. The partitioned feature map is formulated as:

$$X = \left[X_0, X_i, ..., X_{G-1}\right], X_i \in \mathbb{R}^{C//G\times H\times W}, \tag{2}$$

where $C//G$ denotes C divided by G.

Multi-scale spatial information can be collected by large local receptive fields of neurons. Consequently, the EMAM employs three parallel routes to extract attention-weight descriptors. Two parallel routes use 1×1 convolution for smaller receptive fields, while a third path uses 3×3 convolution for a larger receptive field. To capture dependencies across all channels and manage computational budgets, cross-channel information interaction is modeled in the channel direction. Specifically, two 1D global average pooling operations are applied to encode the channel along two spatial directions in the branch with 1×1 convolution. It is formulated as:

$$X_{avg} = \mathrm{avgpool}_x(\mathbf{x}) \in \mathbb{R}^{C//G\times 1\times W}, \tag{3}$$

$$Y_{avg} = \mathrm{avgpool}_y(\mathbf{x}) \in \mathbb{R}^{C//G\times H\times 1}. \tag{4}$$

Meanwhile, a single 3×3 kernel is stacked in the third branch to capture multi-scale feature representation. Reshaping and permuting G groups to align the batch dimension, the EMAM redefines the input tensor with a shape of $C//G \times H \times W$. By employing a similar treatment as Coordinate attention (CA) [20], the EMAM concatenates the two encoded

Authorized licensed use limited to: Queen Mary University of London. Downloaded on November 04,2024 at 15:34:47 UTC from IEEE Xplore. Restrictions apply.

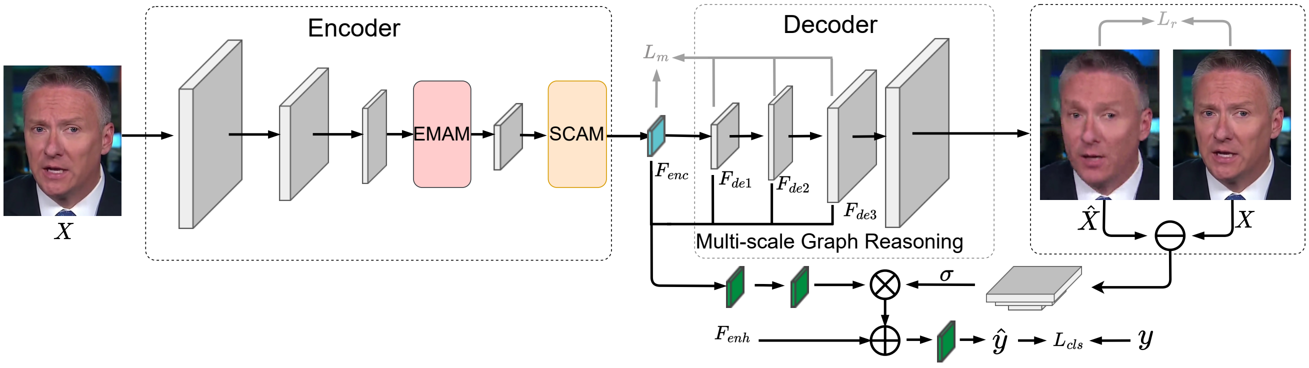

Fig. 1. Schematic diagram of the proposed Progressive Attention Network (PANet).

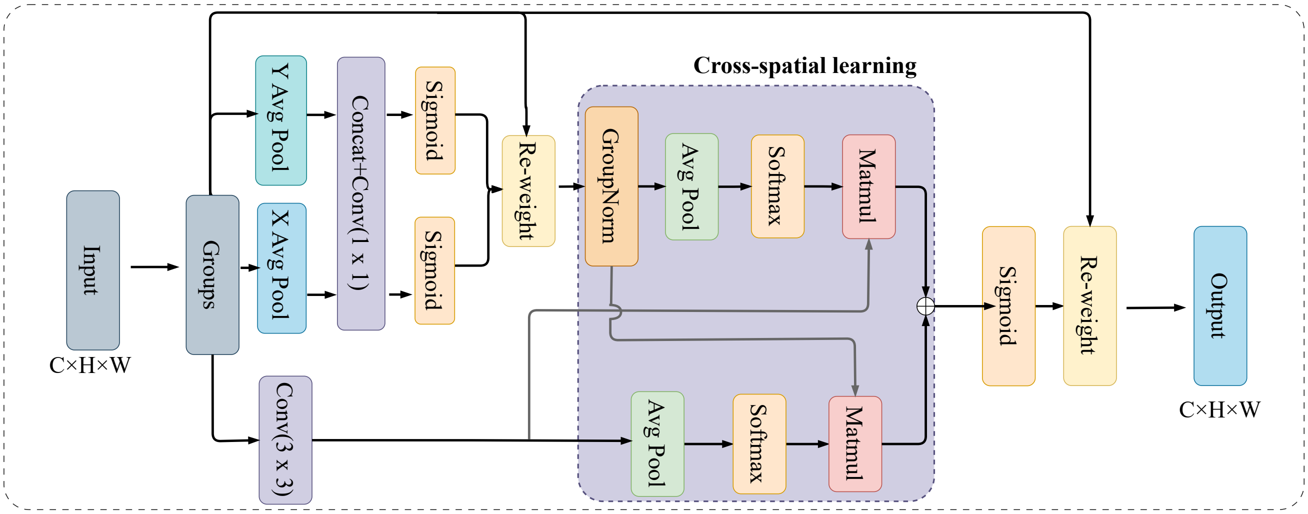

Fig. 2. The structure of the EMAM. The EMAM consists of three parts, namely feature grouping, parallel subnetworks, and cross-spatial learning. Here, "X Avg Pool" represents the 1D horizontal global pooling and "Y Avg Pool" indicates the 1D vertical global pooling, respectively. Cross-spatial learning is employed to aggregate attention-weight descriptors from the parallel subnetworks to capture the pairwise relationships between pixels in the input image.

features along the height direction of the image. It shares the same 1×1 convolution without dimensionality reduction in the first two branches. To model the 2D Binomial distribution over the outputs of the 1×1 convolution, the outputs are factorized into two vectors, and two non-linear sigmoid functions are employed to fit. To establish distinct cross-channel interactive features between the two parallel branches, the two channel-wise attention graphs within each group are amalgamated via simple multiplication. Simultaneously, the third branch captures local cross-channel interaction through a 3×3 convolution to expand the feature space. Consequently, the EMAM incorporates both inter-channel information for modulating the significance of distinct channels and retains detailed spatial structure information within each channel.

The EMAM employs a technique to aggregate information across space in various spatial dimensional directions to achieve more comprehensive feature aggregation. To effectively capture and preserve information, the architecture employs distinct tensors which are the outputs of the 1×1 and 3×3 convolutional branches. EMAM encodes global spatial information in the output of the branch containing 1×1 via 2D global average pooling. Afterward, it directly reshapes the output of the smaller branch to match the corresponding dimensions (i.e., $\mathbb{R}_1^{1\times C//G}\times\mathbb{R}_3^{C//G\times HW}$) before activating the joint activation mechanism for channel features. The formula for the 2D global pooling operation is:

$$y_c=\frac{1}{H\times W}\sum_{j}^{H}\sum_{i}^{W}x_c\left(i,j\right), \tag{5}$$

where x_c indicates the input features at the c-th channel, y_c is the output associated with the c-th channel.

C. Spatial and Channel Attention Module

To achieve fine-grained feature extraction, we introduce spatial and channel attention modules. Further refinement operations were performed to capture fine face forgery traces. The architecture of the spatial and channel attention model is depicted in Fig. 3. The module induces a one-dimensional

Authorized licensed use limited to: Queen Mary University of London. Downloaded on November 04,2024 at 15:34:47 UTC from IEEE Xplore. Restrictions apply.

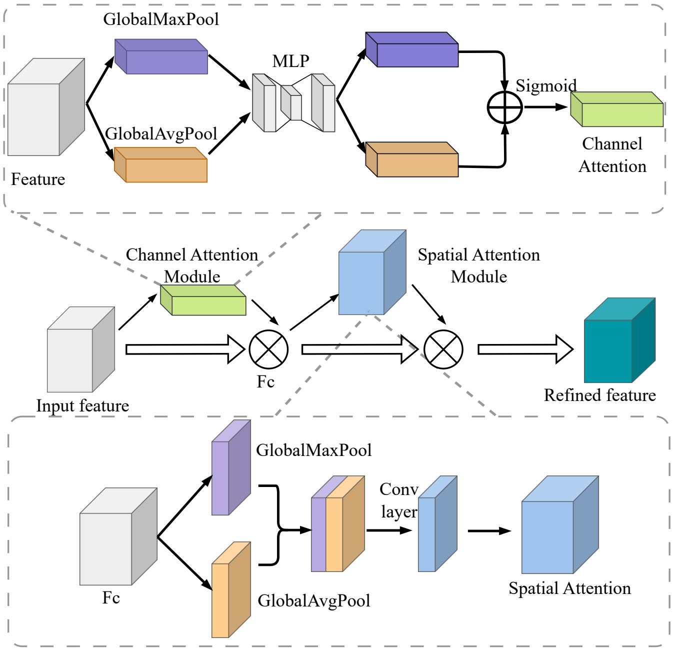

Fig. 3. The structure of the SCAM. The module has two sequential sub-modules: channel attention (CA) module and spatial attention (SA) module.

channel attention mechanism feature map CA and a two-dimensional spatial attention mechanism feature map SA. The computational process is outlined as follows:

$$\begin{aligned} F_c &= CA(F) \otimes F, \\ F_c{}' &= SA(F_c) \otimes F_c, \end{aligned} \tag{6}$$

where F is the input image and $\otimes$ is the element-wise multiplication, The channel attention weighting results in F_c, and the ultimate refined feature map $F_c{}'$ is obtained. The detailed equations for CA and SA are as follows:

$$\begin{aligned} CA &= \sigma\left(f_1\left(f_2(F)\right) + f_1\left(f_3\left(F\right)\right)\right) \\ &= \sigma\left(W_1\left(W_o(F_{avg})\right) + W_1\left(W_o\left(F_{max}\right)\right)\right), \end{aligned} \tag{7}$$

$$SA\left(F\right) = \sigma\left(f\left(f_c\left(F_{avg}, F_{max}\right)\right)\right), \tag{8}$$

where the f_1 stands for Multi-Layer Perceptron (MLP), f_2 and f_3 correspond to the average-pooling and max-pooing functions, respectively. W_0 and W_1 represent the weights of two linear layers. f_c denotes the splicing operation and f denotes the convolution operation. $\sigma(.)$ denotes the sigmoid function. F_{avg} and F_{max} denote the average pooling and the maximum pooling representations, respectively.

D. Loss Function

The loss function consists of three basic elements: reconstruction loss ($\mathcal{L}_\mathrm{r}$), cross-entropy ($\mathcal{L}_\mathrm{cls}$), and metric-learning loss ($\mathcal{L}_\mathrm{m}$). The reconstruction loss $\mathcal{L}_\mathrm{r}$ is denoted as:

$$\mathcal{L}_\mathrm{r} = \frac{1}{|R|}\sum_{i \in R} \|\widehat{\mathbf{X}}_i - \mathbf{X}_i\|_1, \tag{9}$$

where R represents the authentic samples, and $|R|$ denotes the cardinality of R.

Furthermore, a metric-learning loss is employed to minimize the distance between real images and ensure a distinct separation between real and fake images in the embedding space.

$$\mathcal{L}_\mathrm{m} = \frac{1}{N_{RR}}\sum_{i \in R, j \in R} cos(\mathbf{F}_i, \mathbf{F}_j) - \frac{1}{N_{RF}}\sum_{i \in R, j \in F} cos(\mathbf{F}_i, \mathbf{F}_j), \tag{10}$$

where R and F indicate the real sample set and the fake sample set, respectively. $cos(\alpha, \beta)$ is a pairwise distance function calculated using the cosine distance:

$$cos(\alpha, \beta) = \frac{1 - \frac{\alpha}{\|\alpha\|_2} \cdot \frac{\beta}{\|\beta\|_2}}{2}. \tag{11}$$

The overall loss function is defined as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_\mathrm{r} + \lambda_2 \mathcal{L}_\mathrm{m} + \mathcal{L}_\mathrm{cls}, \tag{12}$$

where the weight parameters λ_1 and λ_2 are utilized to balance different losses. $\mathcal{L}_\mathrm{cls}$ is the cross-entropy loss for binary classification.

IV. Experimental results and analysis

A. Experimental Setup

Datasets. The FaceForensics++ dataset (FF++) [21] is a comprehensive and diverse collection designed specifically for deepfake detection. It consists of 1000 real video sequences, each processed using one of four automated face synthesis methods, namely Deepfakes, Face2Face, FaceSwap, and NeuralTextures. In each video sequence, there is a trackable and predominantly unobstructed front face. This contributes to the high realism of the generated faces.

Evaluation Metrics. To evaluate the proposed model, we used the more common evaluation metrics in the binary classification method, *i.e.*, Accuracy (ACC) and Area Under the Receiver Operating Characteristic Curve (AUC). ACC measures the overall performance of the model, while AUC measures the discriminatory power.

Implementation Details. The foundation of the proposed framework lies in the Xception [22] implementation. Training is executed with a batch size of 32, and the epoch is set to 40. A step-learning rate scheduler is implemented for dynamic adjustment of the earning rate. The Adam [23] optimizer was initialized with a learning rate of 2e-4 and a weight decline of 1e-5. In the EMAM feature grouping, it is divided into eight groups. Data augmentation is exclusively achieved through random horizontal flipping. Implemented within the PyTorch framework, the framework leverages the computational power of two 3090 Ti GPUs.

B. Experimental Results

We performed experiments on the benchmark datasets FF++ [21] of different qualities, including high quality (HQ) and low quality (LQ). This ensures the broad applicability and dependability of the experimental results. dependability

Authorized licensed use limited to: Queen Mary University of London. Downloaded on November 04,2024 at 15:34:47 UTC from IEEE Xplore. Restrictions apply.

In each experimental setup, we recorded the key performance indicators, ACC and AUC, to comprehensively evaluate the effectiveness of the proposed method. At the same time, we analyze the data in detail by comparing it with the experimental results of existing methods. By comparison, we aim to highlight the superiority of the PANet over others. The results of the performance of the proposed framework on the FF++ dataset compared to other methods are shown in Table I.

TABLE I
COMPARISON WITH THE SOTA METHODS. THE BEST RESULTS ARE EMPHASIZED IN **BOLD**.

Methods	FF++(HQ)		FF++(LQ)	
	ACC(%)	AUC(%)	ACC(%)	AUC(%)
MesoNet [24]	83.10	-	70.47	-
SPSL [25]	91.50	95.32	81.57	82.82
RFM [26]	95.69	98.79	87.06	89.83
Freq-SCL [27]	96.69	99.28	89.00	92.39
Multi-task [28]	85.65	85.43	81.30	75.59
Face X-ray [29]	-	87.84	-	61.60
Xception [22]	95.73	96.30	86.86	89.30
Add-Net [30]	96.78	97.74	87.50	91.01
Two-branch [31]	96.43	98.70	86.34	86.59
RECCE [8]	97.06	99.32	91.03	**95.02**
PANet(Ours)	**97.26**	**99.37**	**91.37**	94.90

In Table I, the PANet reaches the best results on both ACC and AUC metrics for high-quality (HQ) images. It achieves an ACC of 97.26%, which is a significant improvement of 0.29% over the previous best method. It also achieves an AUC of 99.37%, which is a slight improvement of 0.13% over the previous best method.

On low-quality (LQ) images, the performance gap between the proposed method and other methods is not as significant. However, the PANet still achieves comparable results, with an ACC of 91.37% and an AUC of 94.90%. The PANet achieves the highest ACC and the second-highest AUC among all evaluated methods. Compared with the Xception baseline, the proposed method has achieved an improvement of 2.59% in ACC and 4.23% in AUC. This demonstrates that the proposed method can extract greater discriminative features from low-quality images.

Qualitative results from the high-quality and low-quality subset of the FF++ dataset are shown in Fig. 4 and Fig. 5. In both figures, the top row is the original input images and the bottom row presents the inference results of the proposed network.

V. CONCLUSION

In this paper, we proposed a Progressive Attention Network (PANet) for deepfake detection. The framework incorporated an efficient multi-scale attention module, strategically focusing on crucial facial regions like the nose, mouth, and eyes. Additionally, the spatial and channel attention modules were employed for fine-grained feature extraction. Experimental results show that the PANet reaches ACC and AUC percentages of 97.26% and 99.37%, respectively, for the high-quality subset, and 91.37% for ACC and 94.90% for AUC for the poor-quality subset of the FaceForensics++ dataset. The results proved that the proposed method outperforms state-of-the-art approaches.

ACKNOWLEDGMENT

This work is funded by the project PCI2022-134990-2(MARTINI) of the CHISTERA IV Cofund 2021 program, funded byMCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR; and by the Spanish Ministry of Science and Innovation under FightDIS (PID2020-117263GB-I00) and XAI-Disinfodemics (PLEC2021-007681) grants, and by EMIF managed by the Calouste Gulbenkian Foundation, in the project MuseAI.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[2] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, "Deep learning for deepfakes creation and detection: A survey," *Computer Vision and Image Understanding*, vol. 223, p. 103525, 2022.

[3] B. Wang, X. Wu, Y. Tang, Y. Ma, Z. Shan, and F. Wei, "Frequency domain filtered residual network for deepfake detection," *Mathematics*, vol. 11, no. 4, p. 816, 2023.

[4] H. Huang, N. Sun, and X. Lin, "Blockwise spectral analysis for deepfake detection in high-fidelity videos," in *DSAA*. IEEE, 2022, pp. 1–9.

[5] T. Wang and K. P. Chow, "Noise based deepfake detection via multi-head relative-interaction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, 2023, pp. 14 548–14 556.

[6] L. Chen, Y. Zhang, Y. Song, J. Wang, and L. Liu, "Ost: Improving generalization of deepfake detection via one-shot test-time training," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 597–24 610, 2022.

[7] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *CVPR*, 2021, pp. 2185–2194.

[8] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, "End-to-end reconstruction-classification learning for face forgery detection," in *CVPR*, 2022, pp. 4113–4122.

[9] D. Ouyang, S. He, G. Zhang, M. Luo, H. Guo, J. Zhan, and Z. Huang, "Efficient multi-scale attention module with cross-spatial learning," in *ICASSP*. IEEE, 2023, pp. 1–5.

[10] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018, pp. 3–19.

[11] R. Shao, T. Wu, and Z. Liu, "Robust sequential deepfake detection," *arXiv preprint arXiv:2309.14991*, 2023.

[12] D. Wodajo, S. Atnafu, and Z. Akhtar, "Deepfake video detection using generative convolutional vision transformer," *arXiv preprint arXiv:2307.07036*, 2023.

[13] Y.-J. Heo, W.-H. Yeo, and B.-G. Kim, "Deepfake detection algorithm based on improved vision transformer," *Applied Intelligence*, vol. 53, no. 7, pp. 7512–7527, 2023.

[14] S. Das, S. Seferbekov, A. Datta, M. S. Islam, and M. R. Amin, "Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation," in *ICCV*, 2021, pp. 3776–3785.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018, pp. 7132–7141.

[16] X. Li, X. Hu, and J. Yang, "Spatial group-wise enhance: Improving semantic feature learning in convolutional networks," *arXiv preprint arXiv:1905.09646*, 2019.

[17] J. Zhang, X. Li, J. Li, L. Liu, Z. Xue, B. Zhang, Z. Jiang, T. Huang, Y. Wang, and C. Wang, "Rethinking mobile block for efficient attention-based models," in *ICCV*. IEEE Computer Society, 2023, pp. 1389–1400.

[18] C. Tian, Z. Luo, G. Shi, and S. Li, "Frequency-aware attentional feature fusion for deepfake detection," in *ICASSP*. IEEE, 2023, pp. 1–5.

Authorized licensed use limited to: Queen Mary University of London. Downloaded on November 04,2024 at 15:34:47 UTC from IEEE Xplore. Restrictions apply.

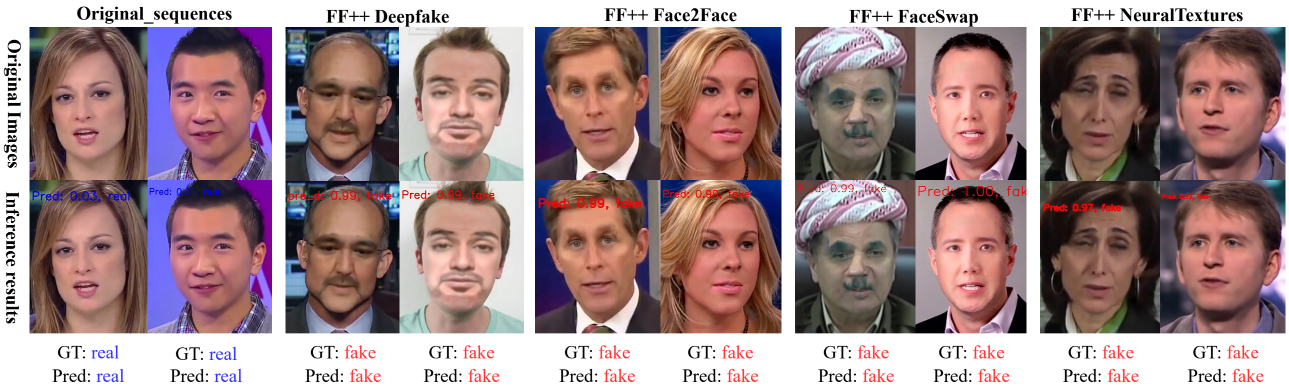

Fig. 4. The qualitative results from the high-quality subset of the FaceForensics++ dataset. The ground truth (GT) indicates the truth label and the prediction (Pred) denotes the predicted label.

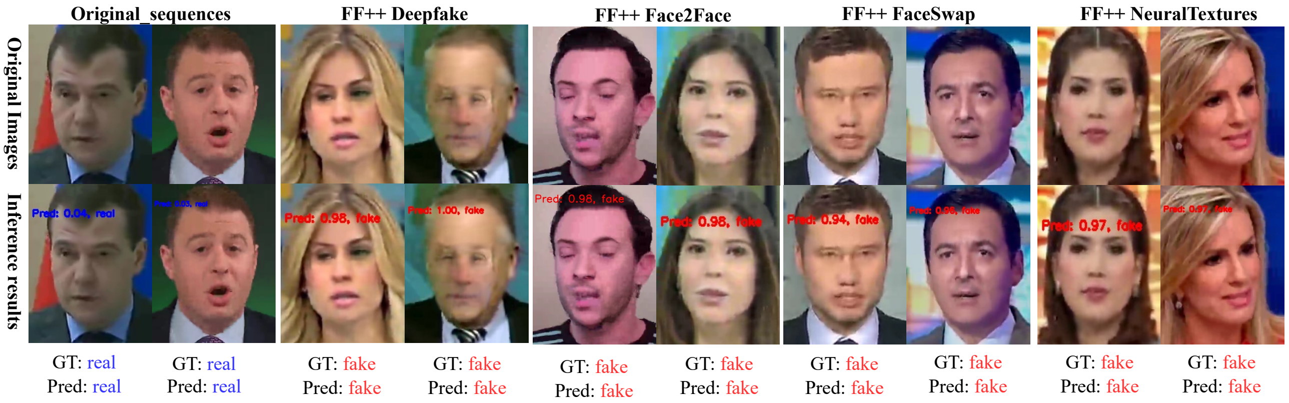

Fig. 5. The qualitative results were extracted from the low-quality subset of the FaceForensics++ dataset. The ground truth (GT) denotes the truth label and the prediction (Pred) denotes the predicted label.

[19] H.-Y. Zhou, C. Lu, S. Yang, X. Han, and Y. Yu, “Preservational learning improves self-supervised medical image models by reconstructing diverse contexts,” in *ICCV*, 2021, pp. 3499–3509.

[20] Q. Hou, D. Zhou, and J. Feng, “Coordinate attention for efficient mobile network design,” in *CVPR*, 2021, pp. 13 713–13 722.

[21] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics++: Learning to detect manipulated facial images,” in *ICCV*, 2019, pp. 1–11.

[22] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *CVPR*, 2017, pp. 1251–1258.

[23] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[24] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “Mesonet: a compact facial video forgery detection network,” in *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2018, pp. 1–7.

[25] 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 *CVPR*, 2021, pp. 772–781.

[26] C. Wang and W. Deng, “Representative forgery mining for fake face detection,” in *CVPR*, 2021, pp. 14 923–14 932.

[27] 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 *CVPR*, 2021, pp. 6458–6467.

[28] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, “Multi-task learning for detecting and segmenting manipulated facial images and videos,” in *BTAS*. IEEE, 2019, pp. 1–8.

[29] 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 *CVPR*, 2020, pp. 5001–5010.

[30] 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.

[31] 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.

Authorized licensed use limited to: Queen Mary University of London. Downloaded on November 04,2024 at 15:34:47 UTC from IEEE Xplore. Restrictions apply.