ShuffleSR: Image Deepfake Detection Using Shuffle Transformer

Cheng Yang
School of Opto-Electronic and Communications Engineering, Xiamen University of Technology
Xiamen, China

Chen Zhu*
School of Opto-Electronic and Communications Engineering, Xiamen University of Technology
Xiamen, China
zhuchen@xmut.edu.cn

Chao Deng
School of Opto-Electronic and Communications Engineering, Xiamen University of Technology
Xiamen, China

Zeyu Xiao
School of Opto-Electronic and Communications Engineering, Xiamen University of Technology
Xiamen, China

2023 IEEE 6th International Conference on Knowledge Innovation and Invention (ICKII) | 979-8-3503-2353-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICKII58656.2023.10332600

***Abstract*—We proposed an improved method based on the Shuffle Transformer for facial forgery detection. The method incorporated SE (Squeeze-and-Excitation) and ROI (Region of Interest) modules. By introducing the SE module, the network's channel attention capability was enhanced and enabled the model to focus more on important feature channels. Meanwhile, the ROI module was employed to extract the region of interest corresponding to the face, enabling the network to concentrate on processing facial regions. The performance of facial forgery detection was improved as a result. Evaluations were conducted on the DFDC dataset and a custom dataset containing various forgery algorithms. The results demonstrated that the enhanced Shuffle Transformer achieved significant performance improvements in facial forgery detection and exhibited better robustness and generalization capabilities while excelling across different forgery algorithms.**

***Keywords*—deep learning, Shuffle Transformer, facial forgery detection, region of interest**

I. Introduction

Face image generation is a popular research topic in image synthesis. The rapid advancements in deep learning, particularly the emergence of Generative Adversarial Networks (GANs) and autoencoders, have empowered existing generation techniques to produce highly realistic and perceptually indistinguishable face images. By amalgamating facial analysis techniques such as face detection and segmentation, specific facial regions can be selectively modified with GAN, thereby expediting the development of facial editing technologies such as face swapping and reconstruction. However, the unauthorized exploitation of facial forgery techniques poses a significant threat to individuals' rights and privacy, reputation, and overall self-image. Consequently, the development of facial forgery detection techniques assumes paramount importance.

The evolution of facial forgery techniques has necessitated the research community to devise effective methodologies for discerning the authenticity of facial images. Concurrently, there has been a surge in the development of transformer models and their variants, tailored to address diverse computational tasks in computer science. Liang proposed the SwinIR model based on the Swin Transformer, which used several residual-connected Swin Transformer blocks (RSTB) as deep feature extraction modules. It showed state-of-the-art performance in experiments related to image super-resolution, image denoising, and JPEG compression artifact removal [1]. On the other hand, Coccomini combined EfficientNet with Vision Transformers by employing EfficientNet as a feature extractor for image patches in the ViT model. This approach achieved comparable performance with only one-third of the parameters[2].

In this study, we proposed an enhanced face forgery detection model based on the shuffle transformer architecture. This novel approach incorporated SE (Squeeze-and-Excitation) blocks and leverages the concept of Region of Interest (ROI) for facial regions. In comparison to conventional convolution neural network (CNN)-based forgery detection models, the proposed model offered several noteworthy advantages: (1) it fostered content-based interactions between image content and attention weights, akin to spatially varying convolutions [3]. (2) The model achieved long-range contextual modeling through the implementation of a sliding window mechanism. (3) Furthermore, it achieved a reduction in parameter complexity while ensuring performance integrity.

II. Related Works

A. Deepfake Generation

The facial forgery techniques are classified into two categories: Generative Adversarial Network (GAN)-based facial image generation techniques and facial editing techniques.

By training the generator and discriminator networks in an adversarial manner, GANs produce highly realistic synthetic facial images. The generator network generates candidate facial images from random noise, which are then evaluated by the discriminator network. As the training progresses, the generator network gradually learns to produce convincing facial images, while the discriminator network enhances its ability to distinguish between real and generated images. Several methods, such as StarGAN and StyleGAN, have been developed to generate forged facial images that are difficult for the human eye to distinguish from genuine ones.

On the other hand, facial editing techniques focus on modifying and manipulating existing facial images. These techniques enable operations such as facial expression changes, age progression, and gender transformation. Generally relying on deep learning models, facial editing techniques leverage the learned facial feature representations and transformation patterns to perform precise editing while preserving the authenticity of the facial content. Several well-

979-8-3503-2353-5/23/$31.00 ©2023 IEEE

known models in this domain include Faceswap, DeepFaceLab, and SimSwap.

B. Deepfake Detection

The detection of forged facial images is broadly categorized into two classes: traditional image processing methods and deep learning methods.

Conventional image processing methods primarily rely on computer vision and image processing techniques to detect forged facial images. These methods typically involve techniques such as feature extraction, texture analysis, and statistical modeling. For instance, methods based on texture analysis such as the Local Binary Pattern (LBP) and Support Vector Machines (SVM) are commonly used.

With the advancement of deep learning technologies, deep neural networks have achieved remarkable success in detecting forged facial images. Deep learning methods leverage large-scale datasets for training and employ architectures such as CNNs, GANs, and Recurrent Neural Networks (RNNs) to extract high-level features from images. These networks are capable of learning richer feature representations, leading to higher accuracy and generalization performance in detecting forged images. For example, the deep learning-based CNN XceptionNet excels in image classification tasks, and the EfficientNet-based method achieves outstanding performance in forgery detection by employing pre-training on large-scale datasets and fine-tuning on forged image data [4].

III. Fundamentals and Network Framework

A. Shuffle Transformer Network Model

In recent years, Transformer-based visual models have achieved remarkable performance in tasks such as image classification, object detection, and semantic segmentation. However, the computational complexity of these models exhibits quadratic growth concerning increasing input image dimensions, thereby limiting their applicability in dense prediction tasks. To address this issue, Tencent GY-Lab proposed a novel Shuffle Transformer model that rethinks the spatial Shuffle operation for establishing connections between windows. The spatial Shuffle operation is used in window-based self-attention modules to establish connections between non-overlapping windows, significantly enhancing the modeling capacity [5]. The model architecture is illustrated in Fig. 1.

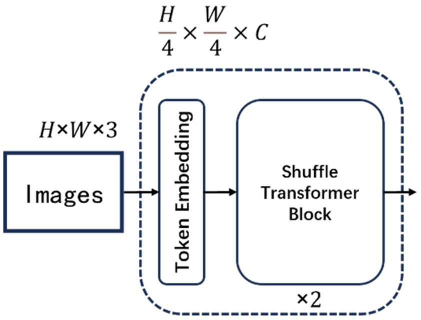
$\frac{H}{4}\times\frac{W}{4}\times C$
H×W×3
Images
Token Embedding
Shuffle Transformer Block
×2

Fig. 1. Architecture of shuffle transformer.

The Shuffle Transformer Block module shown in Fig. 2 uses the Shuffle WMSA for calculating self-attention within a local window based on Window-based Multi-Head Self-Attention (WMSA) mechanism. In the module, an image is evenly divided into non-overlapping windows to perform attention calculations [5]. To improve the connections between adjacent windows, a deep convolutional layer is inserted with a residual connection between the WMSA module and the MLP module. This enhances the information flow between neighboring windows and alleviates the "grid problem" that arises when the image size is much larger than the window size.

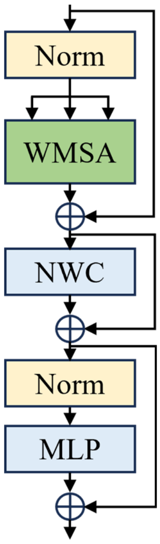

Fig. 2. WMSA shuffle Block

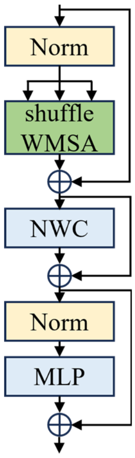

Fig. 3. Shuffle WMSA shuffle Block.

B. Squeeze-and-Excitation block

The primary purpose of the Squeeze-and-Excitation (SE) block is to enable the network to adaptively learn inter-channel relationships of features, thereby capturing correlations and importances among features more effectively [6]. This mechanism is achieved through two key steps: Squeeze and Excitation.

In the Squeeze, the SE block reduces the dimensionality of feature maps by employing Global Average Pooling, resulting in a channel-wise feature descriptor. More precisely, let $z \in R^C$ be the result of global average pooling applied to the feature Uin spatial dimensions $H \times W$, such that each element of z represents (1).

$$z_c = F_{sq}(u_c) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_c(i,j) \tag{1}$$

In the Excitation phase, the SE block includes a fully connected layer to learn the weights for each channel. This fully connected layer takes the channel descriptor vector from the Squeeze phase as input and outputs a weight vector with the same dimension as the number of channels. These weights are then used to perform channel-wise scaling, allowing for important adjustments of different channels' features. The introduction of the SE block enables the network to focus more on channels that are crucial for the current task, reducing attention to irrelevant or redundant channels and enhancing feature representation capability. As a result, the model captures key features from input data, leading to improved generalization ability of the model.

C. Region of Interest block

Region of Interest (ROI) refers to a specific region within an image or video that holds particular significance or attracts interest. In order to concentrate attention and resources on the most relevant and crucial areas and enhance the efficiency and performance of the algorithm, the model's attention was paid to the facial regions within the images, rather than considering the entire image. To locate the facial regions within the images, we employed the Multi-Task Cascaded Convolutional Networks (MTCNN) as a preliminary model to detect key points on the face. Once the facial regions are identified, they are regarded as ROIs. A rectangular bounding box is determined based on the results of the facial detection process, enclosing the location of the face within the image. Subsequently, through feature extraction on the ROI, the ROI is transformed into a suitable feature vector for further processing and subsequent analysis.

D. Improved Shuffle Transformer

The Shuffle Transformer demonstrates strong modeling capability and multi-scale feature integration in image processing tasks. However, in order to better adapt it for facial recognition, specific improvements are needed for the model. Firstly, facial images exhibit variations in scale and pose, necessitating a mechanism to effectively capture these variations. Secondly, different facial regions (such as eyes, mouth, and nose) contribute to facial recognition, thus demanding the network's increased attention to these crucial areas. Based on these requirements, we incorporated the SE module and ROI module into the Shuffle Transformer, forming the SE-R Shuffle Transformer (ShuffleSR). The improved structure of the ShuffleSR block is illustrated in Fig. 4.

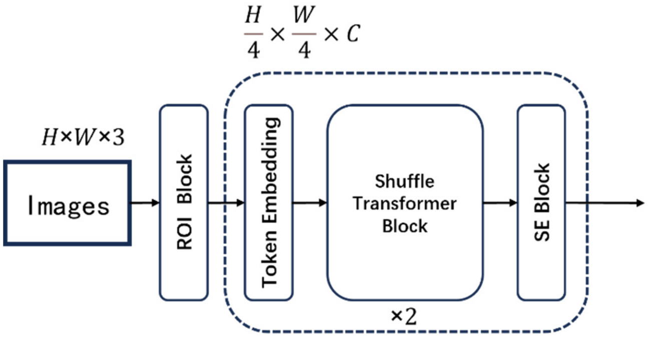
$\frac{H}{4} \times \frac{W}{4} \times C$
$H \times W \times 3$
Images
ROI Block
Token Embedding
Shuffle Transformer Block
SE Block
×2

Fig. 4. Architecture of ShuffleSR

In facial recognition, the SE module assists the network in capturing essential facial features such as eyes and nose, thereby enhancing recognition accuracy. On the other hand, the ROI module is used for localizing and extracting facial regions from the input image, enabling the network to focus on processing these regions specifically. During the training process, emphasizing the facial regions allows the network to better learn and represent facial features, ultimately leading to improved facial recognition performance.

IV. Results and Discussions

A. Experimental Study on DFDC Dataset

The DFDC dataset is a large-scale dataset used for deepfake detection, comprising real and fake face images and videos. The dataset contains challenging and diverse fake videos from various sources and multiple models.

We extracted 120,354 images from the training set of DFDC, using 1,230 images for the validation set. The training set consisted of 63,158 fake images and 57,196 real images. During the training process, we employed the AdamW optimizer with a batch size of 128 and a cosine decay learning rate schedule with an initial learning rate of 0.001 and weight decay of 0.05. All experiments were conducted on a single GTX3090 device. Table I presents a detailed comparison of the ShuffleSR with other image classification networks.

TABLE I. Results on DFDC test dataset

Model	*AUC*	*params*
ViT with distillation [7]	0.974	462M
Efficient ViT [2]	0.921	109M
Convolutional ViT [8]	0.805	89M
ShuffleSR	0.934	88M

In Fig. 5, a detailed ROC plot for the architectures on the DFDC dataset is shown. All models were evaluated using the same number of training and validation samples. The results indicated that the ShuffleSR exhibits superior speed and accuracy.

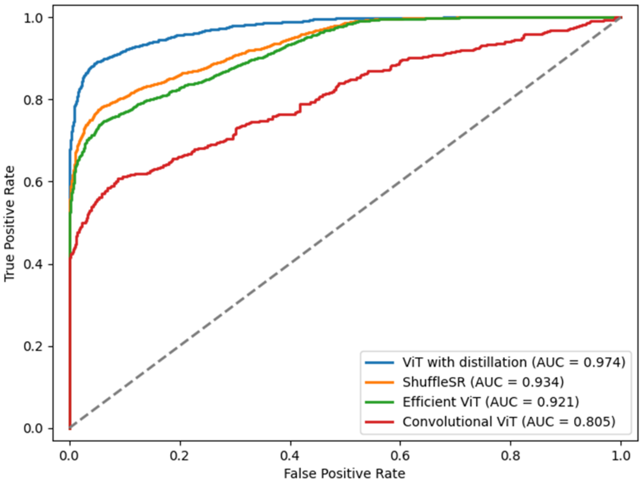

Fig. 5. ROC Curves comparison between our best model and others on DFDC test set.

B. Experimental Study on Data Set Generated by Forgery Algorithms

In this study, we constructed a dataset comprising four forgery algorithms, namely Mean, DeepFakes, FaceSwap, and FaceShifter. Each algorithm generated forged images, labeled as "fake," alongside a collection of genuine facial images labeled as "real." The dataset incorporated forged images sourced from various origins and multiple models, thereby exhibiting challenges and diversities. The genuine images were obtained from authentic facial photographs. This dataset was employed to perform the task of authenticity verification, thereby evaluating the performance of the forgery algorithms we devised in the context of forgery detection tasks.

For each forgery algorithm, we conducted separate experiments, where each experiment consisted of 50,000 forged images and genuine images, ensuring a balanced distribution between the forged and genuine samples. The experimental results are presented in Table II.

TABLE II. EFFICIENTNETV2-S STRUCTURE TABLE

Model	*Mean*	*DeepFakes*	*FaceSwap*	*FaceShifter*
Convolutional ViT[8]	67%	93%	69%	46%
Efficient ViT[2]	76%	83%	78%	76%
ShuffleSR	81%	83%	80%	84%

The experimental results showed that the four models exhibited varying levels of overall accuracy. The ShuffleSR achieved an average accuracy of 81%, outperforming all other models. ShuffleSR performed relatively well under the DeepFakes and FaceSwap forgery algorithms, with accuracies of over 80%. Its best performance with an accuracy of 84%, was observed under the FaceShifter forgery algorithm, showcasing strong recognition capabilities. However, when faced with the Mean forgery algorithm, the model's accuracy was relatively lowered to 81%. This could be attributed to the simplicity of the Mean forgery algorithm, which had a lower degree of adversarial complexity, thereby limiting the full potential of ShuffleSR in this particular scenario.

V. CONCLUSION

We improved to the task of face forgery detection based on the Shuffle Transformer. By introducing SE (Squeeze-and-Excitation) and ROI (Region of Interest) modules, we enhanced the capability to model the importance of channels and regions relevant to the face, enabling the network to focus more on learning and representing facial features. In the experiment, we evaluated the ShuffleSR using the DFDC dataset and a custom dataset containing various forgery algorithms. The results demonstrated the performance of the ShuffleSR in the task of authenticating real and forged faces. Future research is necessary to explore additional feature enhancement methods and attention mechanisms to further enhance the performance and robustness of face forgery detection.

REFERENCES

[1] Liang, J., Cao, J., Sun, G., Zhang, K., van Gool, L., & Timofte, R. (2021). SwinIR: Image restoration using swin transformer. arXiv.

[2] Coccomini, D. A., Messina, N., Gennaro, C., & Falchi, F. (2022). Combining EfficientNet and Vision Transformers for Video Deepfake Detection. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13233 LNCS, 219–229.

[3] Elsayed, G. F., Ramachandran, P., Shlens, J., & Kornblith, S. (2020). Revisiting spatial invariance with low-rank local connectivity. 37th International Conference on Machine Learning, ICML 2020, PartF168147-4, 2848–2859.

[4] Kusniadi, I., & Setyanto, A. (2021). Fake Video Detection using Modified XceptionNet. ICOIACT 2021 - 4th International Conference on Information and Communications Technology: The Role of AI in Health and Social Revolution in Turbulence Era, 104–107.

[5] Huang, Z., Ben, Y., Luo, G., Cheng, P., Yu, G., & Fu, B. (2021). Shuffle transformer: Rethinking spatial shuffle for vision transformer. arXiv.

[6] Hu, J., Shen, L., Albanie, S., Sun, G., & Wu, E. (2017). Squeeze-and-excitation networks. arXiv.

[7] Heo, Y.J., Choi, Y.J., Lee, Y.W., Kim, B.G.: Deepfake detection scheme based on vision transformer and distillation. arXiv preprint arXiv:2104.01353 (2021)

[8] Wodajo, D., Atnafu, S.: Deepfake video detection using convolutional vision transformer. arXiv preprint arXiv:2102.11126 (2021)