EAI Endorsed Transactions
on Pervasive Health and Technology

Research Article **EAI.EU**

A New Deepfake Detection Method Based on Compound Scaling Dual-Stream Attention Network

Shuya Wang[1,*], Chenjun Du[2], Yunfang Chen[2]

[1]Tongda College, Nanjing University of Posts and Telecommunications, Yangzhou, China
[2]College of Computer, Nanjing University of Posts and Telecommunications, Nanjing, China

Abstract

INTRODUCTION: Deepfake technology allows for the overlaying of existing images or videos onto target images or videos. The misuse of this technology has led to increasing complexity in information dissemination on the internet, causing harm to personal and societal public interests.
OBJECTIVES: To reduce the impact and harm of deepfake as much as possible, an efficient deepfake detection method is needed.
METHODS: This paper proposes a deepfake detection method based on a compound scaling dual-stream attention network, which combines a compound scaling module and a dual-stream attention module based on Swin Transformer to detect deepfake videos. In architectural design, we utilize the compound scaling module to extract shallowlevel features from the images and feed them into the deep-level feature extraction layer based on the dual-stream attention module. Finally, the obtained features are passed through a fully connected layer for classification, resulting in the detection outcome.
RESULTS: Experiments on the FF++ dataset demonstrate that the deepfake detection accuracy is 9562%, which shows its superiority to some extent.
CONCLUSION: The method proposed in this paper is feasible and can be used to detect deepfake videos or images.

Received on 11 03 2024; accepted on 16 07 2024; published on 18 07 2024

Keywords: Deepfake detection, compound scaling, channel attention, self-attention, swin transformer

doi:10.4108/eetpht.10.5912

1. Introduction

With the rapid progress of Artificial Intelligence (AI) technology, AI is widely used in various fields such as entertainment, finance, education, tourism, and medical care, bringing great convenience. However, it is also accompanied by the threat of technology abuse, among which deepfake technology [1–5] is particularly prominent. Deepfake is an advanced technology that uses deep learning algorithms to seamlessly replace the face of one person in a video or image with that of another, allowing the target person to appear as if they are saying or doing things that were originally done by the source person. Deepfake techniques are more abused in the fields of extortion, cyber violence, and political struggle, which has a complex impact on information dissemination and society. This threatens national security and public interests. Therefore, the motivation of the paper is to pay attention to and study this technology, finding more effective and innovative ways to deal with the problems and challenges it brings. At present, the mainstream methods to identify and detect deepfakes are by using deep learning technologies [6–12]. Video forgery detection, audio forgery detection, and text forgery detection correspond to three sub-fields of deep learning, namely, computer vision, speech recognition, and natural language processing, respectively.

In the field of video forgery within deepfake technology, this paper proposes a method for detecting deep forgeries based on a Compound Scaling Dual-Stream Attention (CSDSA) network. Swin Transformer, introduced by Liu et al.[13], is a novel image feature

[*]Corresponding author. Email: wangsyz@njupt.edu.cn

extraction network that incorporates a local self-attention mechanism. This makes the transformer structure suitable for processing large-scale images through blocking and cross-grouping, enabling the extraction of multi-sacle feature information at various processing levels. Building upon the Swin Transformer architecture, this paper combines the strengths of Convolutional Neural Networks (CNNs) and Swin Transformer to achieve higher performance in detecting deep forgeries. In summary, the main contributions of the paper are as follows:

- Improvements are made to the downsampling module within the Compound Scaling (CS) approach in the shallow feature extraction layer. By adding a two-dimensional average pooling to the shortcut connection, this modification complets the down-sampling operation in the pooling step and reduces the loss of feature information during the downsampling process;
- The residual channel attention module was incorporated into the deep feature extraction layer to address issues such as gradient vanishing and information loss in deep neural networks. This module allows for the adaptive adjustment of weights for each channel in the network, thereby enhancing the network's feature representation capacity and leading to more stable model training;
- By combining CNN and the Swin Transformer together to absorb the advantages of both, this paper achieves more powerful capabilities for image representation and modeling, resulting in higher performance in deepfake detection.

The structure of the paper is summarized as follows. Section 2 presents related works in deepfake, Section 3 proposes a new architecture for detecting deepfakes in video frames. Section 4 is dedicated to the experimental results and analysis. Section 5 gives a discussion of the experimental results. Finally, conclusions and future work are given in Section 6.

2. Related Works

With the increasing focus on Generative Adversarial Networks (GANs), deepfake and face manipulation techniques have garnered significant attention. These techniques can be categorized into four groups based on the tampering region and purpose: identity swap, face reenactment, attribute manipulation, and entire face synthesis [14–16]. Numerous studies have been conducted in these areas. Below, we highlight some relevant works, and interested readers are encouraged to explore the review literature for more comprehensive coverage [14, 16–18].

(1) Identity Swap

Identity swap involves replacing face photos of a source person with those of a target person in videos or photos, utilizing various algorithms and methods such as GANs, Auto-Encoder (AE), and others. Researchers have proposed methods and technical solutions like FaceSwap [19] and DeepFaceLab [20] to generate high-quality identity swapping forgery images and videos. Currently, GAN-based methods are predominant in identity swap, with CycleGAN [21] being a typical representative work introduced in 2017. Subsequently, many GAN-based methods have been developed, including Faceswap-GAN [22], Face swapping GAN [23], and Region-Separative GAN (RSGAN) [24]. Furthermore, there are several other methods for identity swapping [25–27].

(2) Face Reenactment

Face reenactment is a conditional face synthesis task with dual objectives: transferring the shape of the source face to the target face while preserving the appearance and identity of the target face. Methods for detecting face reenactment primarily rely on CNNs, Recurrent Neural Networks (RNNs), and similar techniques. For instance, Liu et al. [28] proposed a lightweight 3D CNN for deepfake detection, while Kumar et al. [10] proposed the use of multi-stream CNNs to learn region artifacts and achieve robust performance across various compression levels. These methods aim to identify images and videos that have undergone facial reenactment forgery, thus ensuring information security and personal privacy.

(3) Attribute Manipulation

Attribute manipulation involves altering specific attributes (e.g., age, hairstyle) of a face image. This technique is implemented in models like StarGAN [29], AttGAN [30], which utilize GANs to manipulate face attributes and generate realistic tampered images. To detect attribute manipulation, researchers have proposed various methods, including those based on CNNs [31] and color component differences [32]. These methods effectively identify images and videos that have undergone attribute manipulation, thus safeguarding information security and personal privacy. Some notable detection methods include CNNs [33], Efficient-frequency [34], and Attribute Manipulation GAN (AMGAN) [35].

(4) Entire Face Synthesis

Entire face synthesis is a technique that differs from the aforementioned methods in that it generates faces that do not exist in reality, using information such as noise, rather than relying on realistic existing faces.

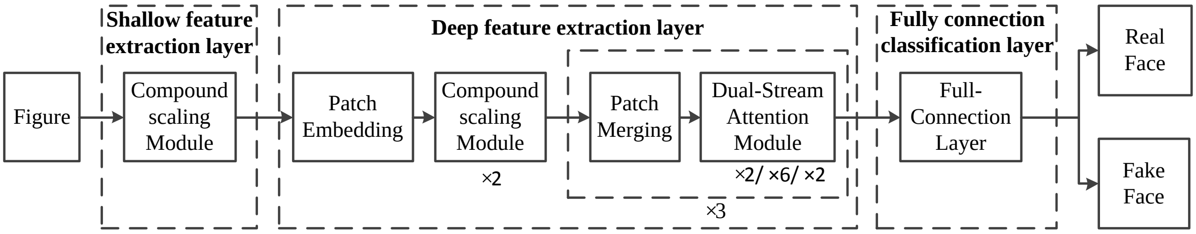


Figure 1. The model architecture of the proposed scheme.

This technique includes methods like Coupled GAN (CoGAN) [36] and Glow [37]. According to existing surveys [10, 28, 38, 39], GANs are the mainstream technique for entire face synthesis. Examples include Progressive Growing GAN (PGGAN) [40] and Style-based GAN (StyleGAN) [41]. Detection techniques for entire face synthesis include methods based on pairwise learning [42] and methods based on image non-uniform response noise [43]. These methods are effective in identifying images and videos that have undergone full-face composite forgery.

3. The Proposed Methodology

This paper proposes a new method for detecting deepfakes, aiming to prevent their abuse. The method is based on a compound scaling dual-stream attention network, as illustrated in Figure 1, which draws on the Swin Transformer structure. The network comprises three layers: shallow feature extraction, deep feature extraction, and fully connected classification. Initially, the compound scaling module is employed to extract shallow feature information from the image. Subsequently, the deep feature extraction layer utilizes the dual-stream attention module to extract deeper features. These extracted features are then forwarded to the fully connected layer for classification, resulting in the detection of deepfake content.

3.1. Extraction of shallow features

Compound scaling method. EfficientNet [44] is a neural network architecture known for its efficiency, achieved through the use of compound scaling to optimize the network's depth, width, and resolution. This approach considers three key dimensions: feature mapping size, the number of channels, and network depth. By applying compound scaling, EfficientNet achieves lightweight models and compression. Different model sizes, such as EfficientNetV2-M and EfficientNetV2-L, are obtained using neural structure search techniques, which involve the use of varying compound scaling coefficients.

In EfficientNetV2, the Mobile Inverted Residual Bottleneck Convolution (denoted as MBConv) module shown in Figure 2(a) is further optimized to the Fused MBConv module, as depicted in Figure 2(b). The primary distinction between MBConv and Fused-MBConv is that in MBConv, the depthwise conv3×3 and expansion conv1×1 are replaced with a single regular conv3×3, leading to improved accuracy and reduced training time of the model. To avoid premature global feature representation in shallow feature extraction, this paper discards the Squeeze-and-Excitation (SE) module, resulting in the CSConv module shown in Figure 2(c), which helps reduce the risk of overfitting.

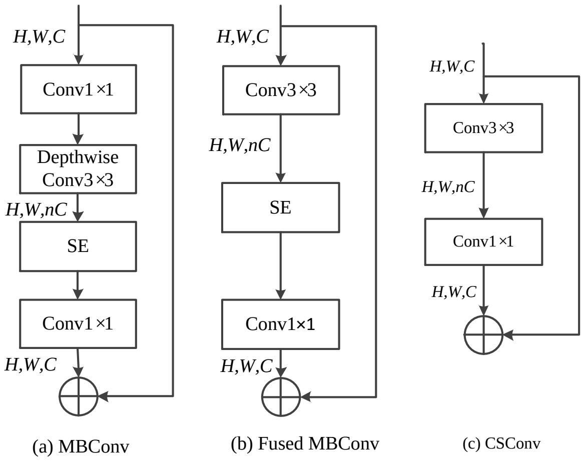
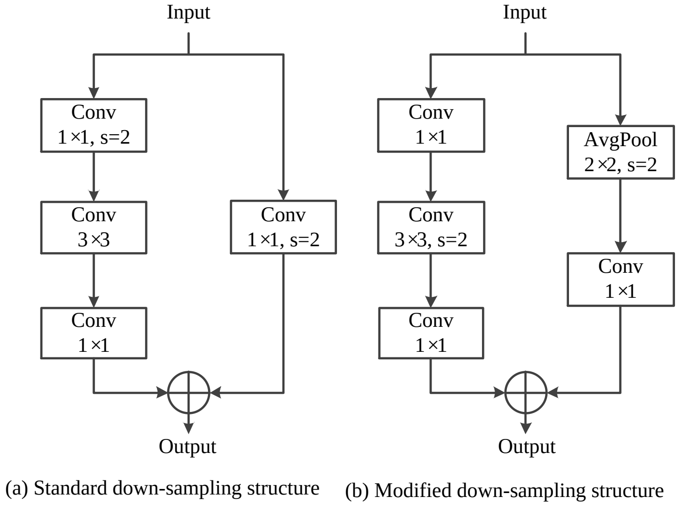


Figure 2. Structure of MBConv, Fused MBConv and CSConv.

Down-sampling method. The standard down-sampling residual structure in Residual Neural Network (ResNet) [45] is illustrated in Figure 3(a). However, this structure has a drawback in the channel dimension: when the stride is set to 2, the 1×1 convolutional operation, instead of the 3×3 convolutional layer, essentially performs a weighted summation along the channel dimension of the input feature maps, leading to the loss of feature information. To address this issue, this paper introduces a modified down-sampling structure called ResNet-D [46], shown in Figure 3(b). ResNet-D adds a two-dimensional average pooling to the shortcut connection, completing the down-sampling operation in the pooling step. This modification helps preserve more feature information during down-sampling.



(a) Standard down-sampling structure (b) Modified down-sampling structure

Figure 3. Modified down-sampling structure.

Compound scaling module. In the three models of EfficientNetV2-L, EfficientNetV2-S and EfficientNetV2-M [47, 48], EfficientNetV2-M ba lances th ree dimensions: accuracy, parameter quantity, and training time. Therefore, a compound scaling module as shown in Figure 4 is proposed based on the partial structure of EfficientNetV2-M by ad ding th e mo dified down-sampling residual structure into it.

In Figure 4, CSConv1 and CSConv4 denote that the amplification f actors o f t he h idden l ayer a re 1 and 4, respectively. The improved down-sampling residual structure replaces the original structure's 3×3 convolution with a stride of 2. Shortcut connections are included to enable gradient propagation backward through the network during training, which alleviates problems such as gradient vanishing and helps the network learn richer feature representations. Additionally, the network can learn multiscale features through these shortcut connections and perform information fusion across different levels.

The core of the compound scaling module is the CSConv, depicted in Figure 2(c), which incorporates two types of convolution operations: fused convolution and squeeze convolution. Fused convolution combines spatial transformation and channel fusion into a single operation, increasing the number of channels in the feature map from the input channels to the hidden channels. The hidden channels are calculated as the input channels multiplied by an expansion factor. This design reduces computational steps, thereby improving computational efficiency. Co mpared wi th MBConv, this approach can reduce the computational cost while maintaining good feature extraction capability. Compression convolution, on the other hand, reduces the number of channels in the feature map from the hidden channels to the output channels using a 1×1 convolutional kernel. This helps the model maintain good performance while keeping a lower parameter count. Fused convolution focuses on convolution operations in the spatial dimension, aiding in capturing local features of the input feature map. Compression convolution, on the other hand, facilitates feature integration between channels, mapping the output feature map of fused convolution to the desired number of output channels. Together, these two operations provide the network with rich feature representations.

3.2. Deep-layer feature extraction method

Residual-channel attention module. Building upon the residual structure and channel attention, the Residual Channel Attention Network [49] introduces the Residual Channel Attention Block (RCAB), illustrated in Figure 5. The block helps prevent gradient vanishing and information loss in deep neural networks, leading to more stable model training. Channel attention enables adaptive adjustment of the weights for each channel in the network, enhancing feature representational capacity and further improving model performance.

The RCAB consists of several key steps: (1) shortcut connections are employed to reduce information loss; (2) image features are extracted using two consecutive 3×3 convolutional layers; (3) the network adjusts the weight for each channel adaptively in the channel dimension using the channel attention mechanism; (4) finally, the shortcut-connected and weighted feature maps are added together to merge the features and produce the output.

Dual-stream attention module. Building upon the Window-based Multi-head Self-Attention/Shifted Window-based Multi-head Self-Attention (W-MSA/SW-MSA) module in Swin Transformer [13] and combined with RCAB, this paper proposes a Dual-Stream Attention (DSA) module, illustrated in Figure 6. The DSA module aims to achieve richer feature representations by integrating the channel attention mechanism from CNN with the self-attention mechanism from Swin Transformer. The RCAB is instrumental in extracting and enhancing important features of specific channels. On the other hand, W-MSA/SW-MSA enables the capture of multi-scale contextual information. The combination of these mechanisms allows for the capture of richer feature representations at various scales and levels, thereby enhancing model performance. Furthermore, W-MSA/SW-MSA is adept at capturing image features at different resolutions by modeling multi-scale information. When combined with RCAB, the model gains strong representation capabilities at different scales, thereby improving its performance

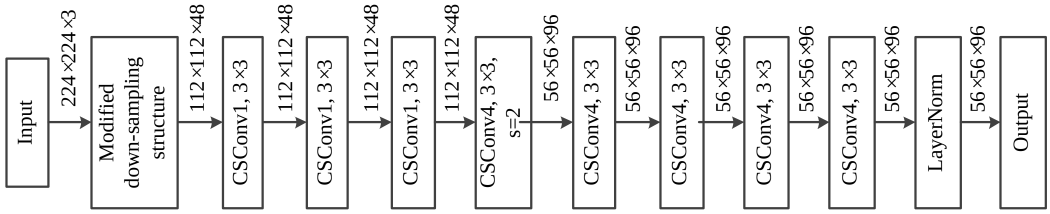


Figure 4. Compound scaling module.

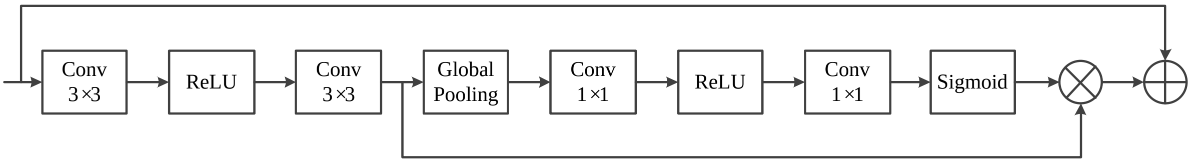


Figure 5. The structure of RCAB.

in multi-scale tasks. In terms of global and local information, RCAB primarily focuses on local features and inter-channel relationships, while W-MSA/SW-MSA can capture long-range dependencies and global information. Combining these mechanisms enables an effective fusion of global and local information, further enhancing the model's performance in complex computer vision tasks.

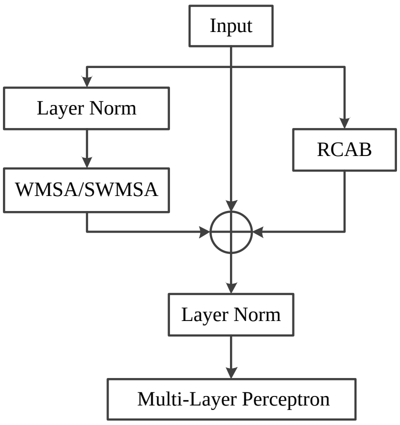


Figure 6. Dual-stream attention module.

4. Experiments and Results

4.1. Data sets and the evaluation metrics

This paper utilizes the FaceForensics++ (FF++) dataset [50] for training, comprising a total of 5000 videos to ensure both quantity and diversity in the training data. In addition to original videos sourced from YouTube, the dataset also includes videos generated by four deep forgery methods: Deepfakes [51], Face2Face [52], FaceSwap [53], and NeuralTextures [54]. The FF++ dataset authors applied varying levels of compression to the real and fake images, resulting in high-quality and low-quality versions of the data, respectively. For this study, the first 50 frames of each video in the high-quality version of the dataset are extracted as the training data. The division ratio for the training set, validation set, and test set follows the official practice of FF++, where the 1000 videos are divided into 720 for training, 140 for validation, and 140 for testing. Table 1 provides statistical information regarding the dataset.

In this paper, accuracy is used as the performance evaluation index, calculated as the ratio of correctly identified examples to the total size of the test set. The calculation formula is as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{1}$$

where TP, TN, FP, and FN represent True Positive, True Negative, False Positive, and False Negative, respectively. In multi-classification problems, the accuracy metric refers to selecting the index corresponding to the maximum value in the probability vector output by the model as the prediction. If this index aligns with the actual class, it indicates a correct prediction.

4.2. Experimental environment and parameter setting

All experiments were conducted on Ubuntu 18, Intel(R) Xeon(R) Platinum 8338C CPU @ 2.60GHz, DDR 320G RAM, four NVIDIA GeForce RTX 3090 (24GB) GPU.

The model hyper-parameter configuration is presented in Table 2, where the channel dimension of the input multi-channel feature extraction layer is reduced from 224 to 96 for the small model or 128 for the base model. Recognizing that as stages progress, the

Table 1. Statistical information of the dataset.

Type	Total Frames	Training Set	Validation Set	Test Set
Original	50k	36k	7k	7k
Deepfakes	50k	36k	7k	7k
Face2Face	50k	36k	7k	7k
FaceSwap	50k	36k	7k	7k
NeuralTextures	50k	36k	7k	7k

Table 2. Model hyper-parameter configuration.

Type	Name	Values of the small model	Values of the base model
Model	Embeding dimension	96	128
	In Channels	3	3
	Num Heads	[3,6,12,24]	[4,8,16,32]
	Window Size	7	7
	Depth	[2,2,6,2]	[2,2,18,2]
Training	Base Learning Rate	5e-5	5e-5
	Epochs	300	300
	Warmup Epochs	20	20
	Warmup Learning Rate	5e-7	5e-7
	Min Learning Rate	5e-6	5e-6
	Decay Rate	0.1	0.1

granularity of feature extraction should become finer, the model proposed in this paper utilizes a compound scaling module for local, multi-level shallow feature extraction on input images. Concurrently, the dual-stream attention mechanism conducts global, multi-scale deep feature extraction on input features, effectively capturing the structure and information of the input data. To balance the channel dimension and the model's depth, the number of self-attention heads in W-MSA/SW-MSA has been adaptively adjusted.

4.3. Experimental results

To ensure the robustness of the results, each experiment was run 20 times, and the average value was taken as the final result. Table 3 presents a comparison of the average accuracies obtained on each subset of the FF++ dataset. From the experimental results of this paper, the deepfake detection model, based on CSDSA network, outperforms the other methods listed in the table in terms of accuracy. Here, CSDSA(S) and CSDSA(B) denote the CSDSA network for the small model and the base model, respectively. It is worth noting that the method described in the FF++ paper involved extracting 100 frames per video, while this paper only extracted 50 frames. In other words, the model proposed in this paper achieves improved accuracy while using only half the amount of data compared to the other methods.

Table 3. Experimental results.

Methods	Accuracy
Steg [55]	70.90%
LD-CNN [56]	78.45%
SB-Conv [57]	82.97%
MesoNet [58]	83.10%
CSDSA(S)	95.60%
CSDSA(B)	95.62%

The results in Table 3 indicate that the small model proposed in this paper achieves a similar accuracy to the base model. Regarding the training iteration process, the small model reached an accuracy of 95.60% at the 190th epoch, while the base model had already reached 95.62% at the 150th epoch. This suggests that the training process can converge faster and approach the local or global optimal solution as the depth of the model increases and the number of model parameters expands. The capacity of the small model may already be well-suited to the complexity of deepfake detection tasks. In this scenario, the deeper base model may not fully leverage its larger capacity advantage, leading to limited performance improvement. Nevertheless, there is still potential for optimizing the accuracy of the base model. For example, one can try adjusting the learning strategy or employing model pruning methods to prevent overfitting.

Table 4. Experimental results of ablation experiments.

Methods	Accuracy of the small model	Accuracy of the base model
CSDSA	95.60%	95.62%
No-CS	91.12%	92.81%
No-DSA	94.66%	94.87%
No-CS DSA	87.04%	91.41%

4.4. Ablation experiments

To evaluate the impact of the CS module and DSA module on the experiments, this paper conducted three categories of ablation experiments: (1) Removal of the CS module, reverting to convolution with a 4×4 kernel and a stride of 4 (denoted as No-CS); (2) Removal of the RCAB structure from the DSA module, reverting to W-MSA/WS-MSA structures (denoted as No-DSA); (3) Simultaneous removal of both the CS module and the RCAB structure from the DSA module (denoted as No-CS DSA). The experimental results are presented in Table 4 under the same experimental conditions and hyper-parameters. The accuracy curve is depicted in Figure 7 and Figure 8.

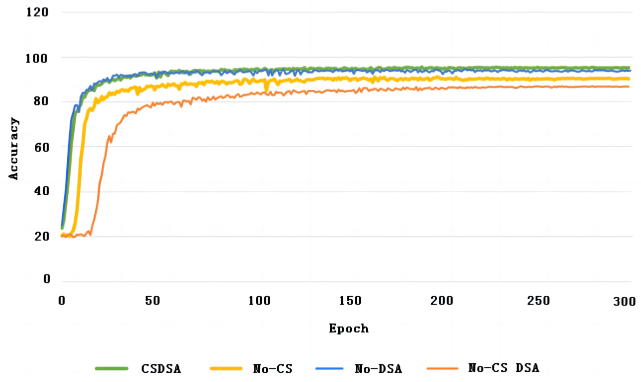


Figure 7. Accuracy curves of the ablation experiments for the small model.

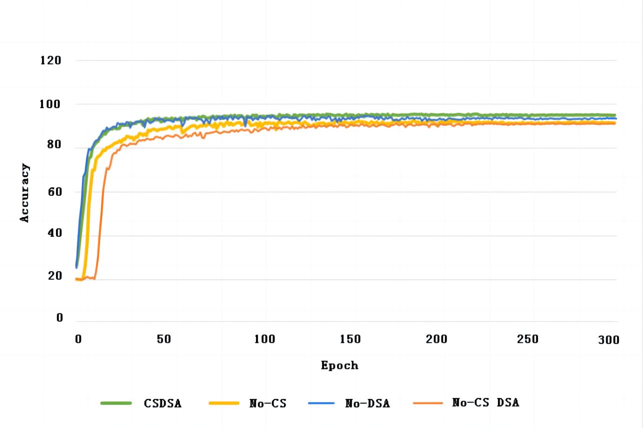


Figure 8. Accuracy curves of the ablation experiments for the base model.

5. Discussion

According to the experimental results of this paper, from the global perspective, the CSDSA model achieved the highest accuracy on both the small and base models, reaching 95.60% and 95.62%, respectively. This success can be primarily attributed to the collaborative efforts of the CS and DSA components, enabling the model to more effectively extract and leverage feature information. Examining the local perspective reveals the contributions of the CS and DSA components to the model performance. Removing the CS module led to accuracy drops to 91.12% and 92.81% for the small and base models, respectively. This suggests that CNN feature extraction plays a crucial role in the model, aiding in capturing local features and texture information within images, thereby enhancing the model's recognition capability. On the other hand, removing the DSA module resulted in accuracy drops to 94.66% and 94.87% for the small and base models, respectively. This indicates that the DSA module, through the introduction of a channel attention mechanism, can adaptively enhance useful features and suppress irrelevant features, further improving the model's performance. When both the CS and DSA modules are removed, the accuracy drops significantly to 87.04% and 91.41%. This further confirms the important contribution of the CS and DSA components to the model performance and their synergistic role in the overall model. In summary, the collaborative and synergistic interaction of key components, including CNN feature extraction in the CS module and the channel attention mechanism in the DSA module, allows the CSDSA model to extract and utilize feature information more effectively, resulting in higher recognition accuracy.

In the ablation experiments for the small model, the epochs at which the individual experiments reached the global or local optimal solutions were as follows: 190 epochs for CSDSA, 165 epochs for No-CS, 213 epochs for No-DSA, and 264 epochs for No-CS DSA. For the large model, the corresponding epochs were 150 for CSDSA, 177 for No-CS, 160 for No-DSA, and 231 for No-CS DSA. With the same hyperparameters and learning rate, the CSDSA scheme converges to the global or local optimal solution faster in the small model. Although the scheme with the CS module removed outperforms the CSDSA scheme in reaching the global or local optimal solution, it is 4.48% less accurate, which is an unacceptable loss of accuracy.

Removing the CS module results in the model lacking a significant number of parameters during the shallow feature extraction process, leading to underfitting, particularly on the FF++ dataset. The large CSDSA model demonstrates a faster convergence towards global or local optimal solutions, further confirming the superiority of the CSDSA approach proposed in this paper on the FF++ dataset. Additionally, although there are some similarities between the accuracy curves of CSDSA and No-DSA, CSDSA not only converges faster in reaching the optimal solution but also achieves the highest accuracy in the end.

The limitations of the proposed system are as follows, and possibly more: (1). Limited Dataset: The system's performance heavily relies on the quality and diversity of the training dataset. A larger and more diverse dataset could potentially improve the system's robustness and generalization capabilities. (2). Performance evaluation indicator: This paper we only consider the accuracy, and temporarily do not consider other performance indicators. In the follow-up work, more performance indicators will be considered for analysis. (3). Robustness to Adversarial Attacks: The system may not be robust against adversarial attacks, where subtle changes to input data can lead to incorrect predictions. Enhancing the system's robustness to such attacks could be an area for future improvement. (4). Limited Scope: The proposed system may be designed for specific types of deepfakes or may not cover all possible variations of deepfake techniques. Its effectiveness against emerging deepfake methods may be limited.

6. Conclusions and Future work

This paper proposes a method to address the issue of local information loss in existing deepfake detection approaches, which combines Swin Transformer and CNN through the design of a CSDSA network. In this approach, the CS module is used to extract shallow local features, optimizing depth, width, and resolution for more efficient local feature extraction. Meanwhile, the DSA module performs deep global feature extraction by combining self-attention and channel attention mechanisms to extract features in both global and channel dimensions. Experimental results demonstrate the superiority of this approach. However, the improvement in accuracy for the base model is less pronounced compared to the small model. This may be due to the small model's capacity already being well-suited to the complexity of deepfake detection tasks, whereas the deeper base model may not fully leverage its larger capacity advantage, resulting in limited performance gains.

Regarding future work, we will continue to optimize the base model to improve its accuracy. This paper only considers accuracy as the performance metric, but other metrics such as detection time, memory usage, etc., can also be considered as research directions. Additionally, other test datasets could be considered in the future to further validate the effectiveness of the proposed method in this paper.

Acknowledgement. This research was funded by the key topics of the "Fourteenth Five-Year" Plan of Educational Science in Jiangsu Province (B/2022/01/174) and the Philosophy and Social Sciences Research Project of Jiangsu Province (2021SJA2438).

References

[1] Nguyen, X.H., Tran, T.S., Nguyen, K.D., et al. Learning spatio-temporal features to detect manipulated facial videos created by the deepfake techniques, *Forensic Science International: Digital Investigation*, 2021, 36: 301108.

[2] Westerlund, M. The emergence of deepfake technology: A review, *Technology innovation management review*, 2019, 9(11): 39-52.

[3] Pantserev, K.A. The malicious use of AI-based deepfake technology as the new threat to psychological security and political stability, *Cyber defence in the age of AI, smart societies and augmented humanity*, 2020: 37-55.

[4] Jones, V.A. Artificial intelligence enabled deepfake technology: the emergence of a new threat, *PhD thesis, Utica College*, 2020.

[5] Neethirajan, S. Is seeing still believing? Leveraging deepfake technology for livestock farming, *Frontiers in Veterinary Science*, 2021, 8: 740253.

[6] Pan, D., Sun, L., Wang, R., et al. Deepfake detection through deep learning, *Proceedings of the 2020 IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT)*, 2020: 134-143.

[7] Deshmukh, A., Wankhade, S.B. Deepfake detection approaches using deep learning: a systematic review, *Lecture Notes in Networks and Systems*, 2020, 146: 293-302.

[8] Chadha, A., Kumar, V., Kashyap, S., et al. Deepfake: an overview, *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security*, 2021: 557-566.

[9] Maksutov, A.A., Morozov, V.O., Lavrenov, A.A., et al. Methods of deepfake detection based on machine learning, *Proceedings of the 2020 IEEE conference of russian young researchers in electrical and electronic engineering*, 2020: 408-411.

[10] Nguyen, T.T., Nguyen, Q.V.H., Nguyen, D.T., et al. Deep learning for deepfakes creation and detection: A survey, *Computer Vision and Image Understanding*, 2022, 223: 103525.

[11] Zhou, L.J., Ma, C., Wang, Z.P., et al. Robust Frame-Level Detection for Deepfake Videos With Lightweight Bayesian Inference Weighting, *IEEE Internet of Things Journal*, 2023, 11(7): 13018-13028.

[12] Yadav, A., Vishwakarma, D.K. AW-MSA: Adaptively weighted multi-scale attentional features for DeepFake detection, *Engineering Applications of Artificial Intelligence*, 2024, 127: 107443.

[13] Liu, Z., Lin, Y., Cao, Y., et al. Swin Transformer: hierarchical vision transformer using shifted windows, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021: 10012-10022.

[14] Juefei-Xu, F., Wang, R., Huang, Y., et al. Countering malicious deepfakes: Survey, battleground, and horizon, *International journal of Computer Vision*, 2022, 130(7): 1678-1734.

[15] Tian, X., Lingyun, Y., Changwei, L., et al. Survey of deep face manipulation and fake detection, *Journal of Tsinghua University (Science and Technology)*, 2023, 63(9): 1350–1365.

[16] Akhtar, Z. Deepfakes Generation and Detection: A Short Survey, *Journal of Imaging*, 2023, 9(1): 18.

[17] Mirsky, Y. and Lee, W. The creation and detection of deepfakes: A survey, *ACM Computing Surveys*, 2021, 54(1): 1-41.

[18] Zhou, X. and Zafarani, R. A survey of fake news: fundamental theories, detection methods, and opportunities, *ACM Computing Surveys*, 2020, 53(5): 1-40.

[19] Korshunova, I., Shi, W., Dambre, J., et al. Fast face-swap using convolutional neural networks, *Proceedings of the IEEE International Conference on Computer Vision*, 2017: 3677–3685.

[20] Liu, K., Perov, I., Gao, D., et al. Deepfacelab: Integrated, flexible and extensible face-swapping framework, *Pattern Recognition*, 2023, 141: 109628.

[21] Zhu, J.Y., Park, T., Isola, P., et al. Unpaired image-to-image translation using cycle-consistent adversarial networks, *Proceedings of the IEEE International Conference on Computer Vision*, 2017: 2223–2232.

[22] Lin, B.S., Hsu, D.W., Shen, C.H., et al. Using fully connected and convolutional net for GAN-based face swapping, *Proceedings of the 2020 IEEE Asia Pacific Conference on Circuits and Systems*, 2020: 185–188.

[23] Nirkin, Y., Keller, Y., Hassner, T. Fsgan: Subject agnostic face swapping and reenactment, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019: 7184–7193.

[24] Natsume, R., Yatagawa, T., Morishima, S. Rsgan: face swapping and editing using face and hair representation in latent spaces, *Special Interest Group on Computer Graphics and Interactive Techniques Conference*, 2018: 1–2.

[25] Zhou, H., Liu, Y., Liu, Z., et al. Talking face generation by adversarially disentangled audio-visual representation, *Proceedings of the AAAI conference on Artificial Intelligence*, 2019: 9299–9306.

[26] Li, L., Bao, J., Yang, H., et al. Advancing high fidelity identity swapping for forgery detection, *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2020: 5074–5083.

[27] Chen, R., Chen, X., Ni, B., et al. Simswap: An efficient framework for high fidelity face swapping, *Proceedings of the 28th ACM International Conference on Multimedia*, 2020: 2003–2011.

[28] Verdoliva, L. Media forensics and deepfakes: an overview, *IEEE Journal of Selected Topics in Signal Processing*, 2020, 14(5): 910–932.

[29] Choi, Y., Choi, M., Kim, M., et al. Stargan: unified generative adversarial networks for multi-domain image-to-image translation, *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018: 8789–8797.

[30] He, Z., Zuo, W., Kan, M., et al. Attgan: Facial attribute editing by only changing what you want, *IEEE transactions on image processing*, 2019, 28(11): 5464–5478.

[31] Marra, F., Gragnaniello, D., Cozzolino, D., et al. Detection of gan-generated fake images over social networks, *Proceedings of the 2018 IEEE conference on multimedia information processing and retrieval*, 2018: 384–389.

[32] Li, H., Li, B., Tan, S., et al. Detection of deep network generated images using disparities in color components, *arXiv preprint*, 2018: 1–26.

[33] Akhtar, Z., Mouree, M.R., Dasgupta, D. Utility of deep learning features for facial attributes manipulation detection, *Proceedings of the 2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence*, 2020: 55–60.

[34] Du, C.X.T., Trung, H.T., Tam, P.M. Efficient-frequency: a hybrid visual forensic framework for facial forgery detection, *Proceedings of the 2020 IEEE symposium series on Computational Intelligencee*, 2020: 707–712.

[35] Ak, K.E., Lim, J.H., Tham, J.Y., et al. Efficient-frequency: a hybrid visual forensic framework for facial forgery detection, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019: 10541–10550.

[36] Liu, M.Y. and Tuzel, O. Coupled generative adversarial networks, *Advances in neural information processing systems*, 2016, 29: 1-9.

[37] Kingma, D.P. and Dhariwal, P. Glow: generative flow with invertible 1×1 convolutions, *Advances in neural information processing systems*, 2018, 31: 1-10.

[38] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., et al. Deepfakes and beyond: A survey of face manipulation and fake detection, *Information Fusion*, 2020, 64: 131–148.

[39] Lyu, S. Deepfake detection: Current challenges and next steps, *Proceedings of the 2020 IEEE international conference on multimedia & expo workshops*, 2020: 1–6.

[40] Karras, T., Aila, T., Laine, S., et al. Progressive Growing of GANs for Improved Quality, Stability, and Variation, *Proceedings of the International Conference on Learning Representations*, 2018: 1–26.

[41] Karras, T., Laine, S., Aila, T. A style-based generator architecture for generative adversarial networks, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 4401–4410.

[42] Hsu, C.C., Zhuang, Y.X., Lee, C.Y. Deep fake image detection based on pairwise learning, *Applied Sciences*, 2020, 10(1): 370.

[43] Marra, F., Gragnaniello, D., Verdoliva, L. Do gans leave artificial fingerprints?, *Proceedings of the 2019 IEEE conference on multimedia information processing and retrieval*, 2019: 506–511.

[44] Tan, M. and Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks, *Proceedings of the International Conference on Machine Learning*, 2019: 6105–6114.

[45] He, K., Zhang, X., Ren, S., et al. Deep residual learning for image recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 770–778.

[46] He, T., Zhang, Z., Zhang, H., et al. Bag of tricks for image classification with convolutional neural networks, *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, 2019: 558–567.

[47] Tan, D.X., Le, Q. EfficientNetV2: Smaller models and faster training, *International conference on machine learning*, 2021: 10096–10106.

[48] Liang, S., Liu, R.H. and Qian, J.S. Fast saliency prediction based on multi-channels activation optimization, *Journal of Visual Communication and Image Representation*, 2023, 94: 103831.

[49] Wang, F., Jiang, M., Qian, C., et al. Residual attention network for image classification, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 3156–3164.

[50] Rossler, A., Cozzolino, D., Verdoliva, L., et al. Faceforensics++: Learning to detect manipulated facial images, *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019: 1–11.

[51] https://github.com/deepfakes/faceswap.

[52] Thies, J., Zollhofer, M., Stamminger, M., et al. Face2face: Real-time face capture and reenactment of rgb videos, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2387–2395.

[53] https://github.com/MarekKowalski/FaceSwap/.

[54] Thies, J., Zollhöfer, M., Nießner, M. Deferred neural rendering: image synthesis using neural textures, *Acm Transactions on Graphics*, 2019, 38(4): 1–12.

[55] Fridrich, J. and Kodovsky, J. Rich models for steganalysis of digital images, *IEEE Transactions on information Forensics and Security*, 2012, 7(3): 868–882.

[56] Fridrich, J. and Kodovsky, J. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection, *Proceedings of the 5th ACM workshop on information hiding and multimedia security*, 2017: 159–164.

[57] Bayar, B. and Stamm, M.C. A deep learning approach to universal image manipulation detection using a new convolutional layer, *Proceedings of the 4th ACM workshop on information hiding and multimedia security*, 2016: 5–10.

[58] Afchar, D., Nozick, V., Yamagishi, J., et al. Mesonet: a compact facial video forgery detection network', *Proceedings of the 2018 IEEE international workshop on information forensics and security*, 2018: 1–7.