Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 21(3): 4142–4164.
DOI: 10.3934/mbe.2024183
Received: 20 December 2023
Revised: 16 February 2024
Accepted: 18 February 2024
Published: 26 February 2024

Research article

Transformer-based cascade networks with spatial and channel reconstruction convolution for deepfake detection

Xue Li, Huibo Zhou*and Ming Zhao

School of Mathematical Sciences, Harbin Normal University, Harbin 150025, China

* **Correspondence:** Email: zhouhuibo@hrbnu.edu.cn.

Abstract: The threat posed by forged video technology has gradually grown to include individuals, society, and the nation. The technology behind fake videos is getting more advanced and modern. Fake videos are appearing everywhere on the internet. Consequently, addressing the challenge posed by frequent updates in various deepfake detection models is imperative. The substantial volume of data essential for their training adds to this urgency. For the deepfake detection problem, we suggest a cascade network based on spatial and channel reconstruction convolution (SCConv) and vision transformer. Our network model's front portion, which uses SCConv and regular convolution to detect fake videos in conjunction with vision transformer, comprises these two types of convolution. We enhance the feed-forward layer of the vision transformer, which can increase detection accuracy while lowering the model's computing burden. We processed the dataset by splitting frames and extracting faces to obtain many images of real and fake faces. Examinations conducted on the DFDC, FaceForensics++, and Celeb-DF datasets resulted in accuracies of 87.92, 99.23 and 99.98%, respectively. Finally, the video was tested for authenticity and good results were obtained, including excellent visualization results. Numerous studies also confirm the efficacy of the model presented in this study.

Keywords: deepfake detection; SCConv; transformer; redundant; visualization

1. Introduction

A fake video of Ukrainian President Zelensky declaring surrender on March 16 during the 2022 Russian-Ukrainian battle went viral on social media. This caused a frenzy among netizens. As intelligent information technology becomes more widespread, people's lives increasingly intertwine with social media and short videos. Deepfake video technology [1], which was initially used in art to create a new video by swapping out the images in the original video, now holds a significant position in the new era. Currently, people can create hyper-realistic fake videos. They use off-the-shelf models with minimal effort or expertise. This raises the possibility of spreading false media messages. It could even be used

to disrupt the societal system. However, this technology has also brought about many negative impacts.

Previous researchers have proposed a variety of related algorithms and detection performance research for deepfake video detection, including those based on biological [2] data, image null-domain features [3], the temporal characteristics of video frames, image frequency-domain features, and more. According to [4] and others, the fabricated video will blur the lips between adjacent frames. This happens because the human visual system can detect minor irregularities in the movement of the created mouth. The network will pick up detailed spatiotemporal representations of the mouth cavity. By applying the learned skills to face deepfake detection, Yu et al. [5] presented the lip-forensics model. It is suggested that an attribution network design be used. The model makes use of the GAN fingerprint data that was obtained to find the fake videos. Although there has been significant progress in studying convolutional neural network-based architectures, there is still room to enhance their accuracy and computation. This paper centers on analyzing videos generated through face swapping and face reproduction, aiming to advance the efficacy of deepfake video detection. To this end, we introduce an enhanced convolutional visual transformer. Convolutional neural networks (CNN) [6] can recognize and characterize fake facial features. This mechanism is employed to extract distinctive features from the manipulated facial images. CNN is essential to identify the irregularities in the picture patches. It is cascaded with a transformer [7] and used as a feature extractor. As a result, we decided to use the enhanced CNN and vision transformer (ViT) as parts of the deepfake detector. It is called transformer-based cascade networks with spatial and channel reconstruction convolution (SCViT). We anticipate great success from the union of these two elements.

CNN is employed to extract and comprehend visual features. The ViT assimilates more meaningful correlations from the input sequence of information to enhance detection efficacy. The ensuing elucidation encompasses the principal contributions emanating from our research:

1) We present an enhanced convolutional vision converter model. It utilizes a conventional convolutional neural network (CNN) as a feature extractor. The model employs spatial and channel reconstruction convolutional techniques to reduce feature redundancy [8]. The next classification step uses the extracted features as input, which can increase detection precision while requiring less processing and input space.

2) To enhance the transformer encoder's versatility in managing arbitrary size, we introduce zero-padding positional coding into the feed-forward network of the ViT.

3) Through comparison tests, we confirm the efficacy of our technique. This confirmation is done at both the picture and video levels. We use many deepfake datasets, deepfake videos, preprocessing, and data augmentation.

The structural organization of this manuscript is outlined as follows: Section 2 summarizes the related work and describes the fake video techniques and the fake video detection methods used in the study. Section 3 details the network architecture proposed in this paper and outlines the dataset used for the study. Section 4 focuses on the procedural aspects of the experimental design, the subsequent experimental results, and the visualization results. Section 5 concludes and looks forward to future work.

2. Related work

Hyper-realistic deepfake images [9], videos [10], and audio signals can be produced with ease thanks to the rapid development of cellular neural networks [11] and generative adversarial networks (GANs).

We analyze the current deepfake techniques and the deepfake detection methods currently put forth by other researchers.

2.1. Deepfake techniques

Currently, deepfake generation techniques are rapidly becoming more sophisticated. Popular techniques include encoder-decoder, CNN, and generative adversarial networks. Mukhopadhyay et al. [12] proposed LipGAN toward automatic face-to-face translation. A proposed method based on GAN technology can provide realistic lip synchronization between a person's native language and the target language in video form. Prajwal et al. [13] proposed that the Wav2Lip model uses a trained lip-sound synchronization model to oversee the model and produce natural-sounding, human-like speech. With precise lip motion, the "master and puppet" connection is likened to a puppet master. Deepfake detection can also be categorized as face replacement, facial reenactment, attribute editing, etc., for different facial regions, as well as the difference in the goal of the tampering, as shown in Figure 1. This technique is frequently used in movie post-production to dub actors' voices or alter their facial expressions.

Figure 1. Examples of face synthesis using three different deepfake techniques.

2.1.1. Face-swapping

Face-swapping technology, characterized by the transplantation of one individual's facial features onto another person's visage, is another name for face replacement. Face swapping [14] employs a 3D face reconstruction model. The final target image is obtained by using a 3D face reconstruction model to model the essential features of the face, render the texture of the 3D face model, and then apply affine transformation, color correction, and other procedures to the image—an occlusion-aware, high-fidelity face swap algorithm with two stages. Xu et al. [15] presented a lightweight identity-aware dynamic network (IDN) for face swapping by dynamically changing model parameters based on the identification information. Wang et al. [16] introduced a practical attribute-preserving framework known as AP-Swap. This framework utilizes a landmark-guided feature entanglement module and a global residual attribute-preserving encoder. Its purpose is to perform face swapping while significantly preserving important facial attributes.

2.1.2. Reenacting a face

With the help of facial reenactment [17], we can alter the head stance and facial expression of one individual while maintaining their identity. These algorithms, specifically, the process involves using a source face image as input. Various strategies are implemented to ensure changes in traits related to facial expressions. The goal is to prevent any alteration of identifying information. As an illustration, some researchers [18] have used 12-dimensional gesture coding to control the position of the face, giving the video character a genuine appearance that makes it appear as though they are changing their expressions. Furthermore, several researchers [19] have developed the concept of a neural radiation field. This field directly considers the characteristics of the audio stream to create a dynamic neural radiation field. It provides a more detailed depiction of the facial video, resulting in a more realistic appearance. To produce more realistic synthetic effects, several researchers [20] focus on creating more natural, smooth, and vibrant face films by separating the speech content from the speaker's identifying information in the audio. A neural face-reenacting method called HyperReenact [21] has been proposed. It serves as a solution to the problem of artifacts generated during face reenacting. Additionally, face keypoint estimation modules [22] are utilized to generate highly realistic reenactment videos of conversations. In conclusion, facial reenactment algorithms have numerous potential applications in various fields. They excel at creating synthetic face videos that appear highly genuine and natural while preserving the identity of the target person.

2.1.3. Altering attributes

The process of altering a specific region of an attribute, such as skinning the face, erasing scars, or adding eyeglasses, is known as attribute manipulation. This technique can also involve more powerful features that change attributes like gender and age without affecting other regions of the same attribute. A hybrid scattering-based module, named WS-SE [23], is introduced for face attribute classification. This module integrates frequency-domain (WST) and image-domain features in a channel-attention manner. To improve the interaction between space generative adversarial networks (GANs), Xu et al. [24] adopted a strategy based on the transformer architecture. Additionally, Sun et al. [25] proposed a 3D perceptual generator based on neural radial fields, which enhances the consistency of the generated images between various points of view by spatially aligning semantic, geometric, and textural information. This makes the results of attribute manipulation highly effective in quality and increases the generalization.

2.2. Deepfake detection

The network architecture examined in this paper is based on image null domain features, and various architectures designed and developed for deepfake detection are discussed below. Deepfake detection methods can be categorized according to biological information, temporal features of video frames, image frequency domain features, and image null domain features.

Since these models are not designed explicitly for deepfake detection, they can be used for various image classification tasks, including deepfake detection. Examples of such models include ResNet [26], DenseNet [27], Xception [28] and EfficientNet [29]. Insufficient generalization for deepfake films produced by various fraud detection systems is evident in this model. It lacks specialization in deepfake detection. Afchar combines the inception module with MesoNet [30], a lightweight CNN-based network

that focuses on the deepfake techniques Face2Face and DeepFake, and Zhao et al. [7], who contend that since there is little distinction between real and fake faces for deepfake detection, the model should be generalized to include both types of faces. Deepfake detection is considered a fine-grained classification challenge since minute distinctions exist between actual and phony faces. They introduced a module for texture enhancement and another for attention-getting. A bilinear attention pooling module was also implemented to guide the network's attention to different localizations. This process magnifies subtle artifacts in shallow features, enhancing the model's accuracy in face feature extraction and increasing the effectiveness of deepfake detection.

Zhao et al. [31] introduced pairwise self-consistency learning (PCL) by measuring the consistency of the source features in the image and calculating the cosine similarity of the local blocks between two grayscale maps. The method aims to explore an efficient and dependable representation for deepfake detection. By dynamically creating annotations of forged images and their operational zones, inconsistent image generators are also suggested as an effective way to help PCL training. Shiohara et al. [32] proposed self-blended images (SBI). This method synthesizes a forgery image by fusing the fake source and target images from a single original image. This approach enhances the model's generalization of unknown manipulations and scenes. This method exhibits excellent generalization capacity. It is effective for fake photographs created by unidentified forgery methods. The model does not rely on the fake faces of a specific forgery method for training. [33] introduce conditional decoder and contrast regularization loss so the model avoids overfitting. Xu et al. [34] proposed the multi-channel xception attentional pairwise interaction (MCX-API) approach. This approach uses pairwise learning and data from various color space representations. The goal is to capture complementary information in a fine-grained manner. This method transforms the underlying network into a multi-channel network. It executes pairwise learning by mimicking attentional pairwise learning. The approach is believed to apply to new methods of forgery creation. Two further losses [35] are introduced to allow the CNN backbone to integrate face images into the implicit identity space: the explicit identity comparison (EIC) loss and the implicit identity exploration (IIE) loss. A face forgery detection framework [36] with multi-scale adapters is proposed based on SAM. A reconstruction-guided attention (RGA) module is also introduced to enhance generalization. [37] proposed the CSTD method to fully utilize spatio-temporal inconsistent information through a two-stage spatio-temporal video encoder. A fine-grained spatial frequency distillation module enhances the detection of high-compression depth forged video. Simultaneously, a mutual information temporal contrast extraction module is introduced for effective detection.

3. Method

We suggest methods for detecting deep forgeries in this section. To detect fake videos, our network model's feature extraction section uses ordinary convolution and SCConv [8]. We also enhanced ViT's [38] feed-forward layer, and the data preprocessing section includes face extraction and data enhancement. Section 3.1 describes the function and characteristics of SCConv in deepfake detection. Section 3.2 gives our entire network architecture and the data preprocessing procedure.

3.1. Data preprocessing

Deepfake datasets are expanding in variety because of the constant upgrading of deepfake creation algorithms, moving beyond the early days of limited data volumes, murky films, the development of

supplying audio as well as the appearance of several faces in the same film, and video homogeneity forging faces with more excellent quality and more lifelike. Figure 2 shows a sample of the extracted faces.

Figure 2. Extracted datasets comprising real and fake images are presented, with the Celeb-DF-v2 dataset on the left and the DFDC dataset on the right.

We primarily utilize the following openly accessible datasets for this paper:

1) FaceForensiac++ [39] dataset: In 2019, Rössler et al. published FaceForensiac++, a well-known dataset in falsified videos. The dataset comprises 5000 synthetic films generated using fake techniques such as DeepFakes, Face2Face, FaceSwap, NeuralTextures and 1000 real videos. Despite being traditional and famous, it has subpar video quality and blatantly fake features.

2) DFDC [40] dataset: one of the most challenging deepfake detection datasets currently available, was proposed by Dolhansky et al. The dataset comprises 19,154 authentic videos and 99,992 manipulated videos; the original videos are all filmed by actors; the videos contain many interferences such as compression, extreme lighting, etc.

3) Celeb-DF [41] dataset: consists of 5626 fake videos and 590 actual videos. The increased data volume and better video resolution dramatically lessen artifacts in the fabricated videos when using an upgraded deepfake face technique.

Table 1. Information pertaining to datasets utilized for deepfake analysis.

Datasets	Original videos	Fake videos	Source
FaceForensics++	1000	4000	Youtube
DFDC	19,154	99,992	Deepfakes competition
Celeb-DF	590	5962	Youtube

The Celeb-DF-v2 dataset has dramatically expanded the earlier Celeb-DF-v1 dataset. In Table 1, Celeb-DF videos exhibit relatively high resolution and fewer artifacts in the manipulated videos. In

contrast, the DFDC dataset comprises 19,154 authentic videos and 99,992 abused videos, presently acknowledged as one of the most formidable datasets within the domain of deepfake detection. The FaceForensics++ dataset is widely used and contains four deepfake methods for generating fake videos using Face2Face, FaceSwap, DeepFakes, and NeuralTextures. We chose these three datasets as our base dataset. After selecting the dataset, we used BlazeFace, a sophisticated face detector, to preprocess the dataset to identify and extract the face region from each frame of images.

3.2. The proposed method

There are two primary steps to our suggested bogus video-detecting system. We start by preparing the data. In this stage, the forged video is split into several frame pictures via frame-splitting. Next, we find and extract face regions from each frame image using Blaze-Face [42], a sophisticated face detector. The retrieved face photos are then cropped to create consistently sized face photographs. These pre-processed photos are then fed into our phony video detector. Our detection model, SCViT, comprises a ViT-based module incorporating SCConv for video authenticity assessment and a feature learning component dedicated to acquiring salient features from the input image. After the detection, the face detection frames and authenticity detection results are tagged to each frame. Finally, new video output is synthesized, thus completing the detection process by displaying the detection results of the video frame by frame. We may examine every frame in this method to see if it contains a deepfake. The secret to this technology is that it can accurately identify forged videos and deliver dependable authenticity judgment for each frame thanks to rigorous data preparation and effective detection models.

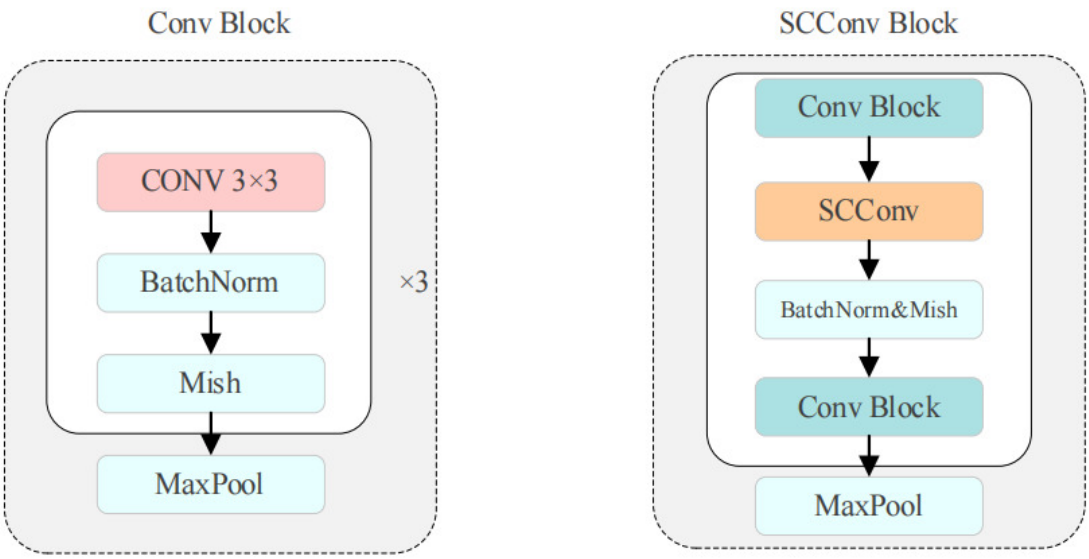

Figure 3. Two convolutional modules in the backbone network.

3.2.1. Overall network structure

SCConv and ViT components make up our SCViT model. The image feature map produced by the SCConv component serves as the input to the ViT. By primarily using convolutional techniques

and the ViT architecture, we enhanced the original ViT. To improve efficiency and decrease model parameters and floating point operations (FLOPs), we used a stack of SCConv and regular convolutions for effective feature extraction. We employed 16 standard convolutions and 1 SCConv in the model's front section, with the SCConv coming after the first standard convolutional layer. Our convolutional modules are stacked as in Figure 3. To safeguard against inter-layer modifications that might impede the learning dynamics of the CNN architecture, and the convolution kernel is applied with a step size of 1 and padding of 1. Following each standard convolution, batch normalization and the Mish activation function [43] procedures, known for their exemplary generalization and optimization capabilities, are implemented.

Furthermore, a maximum pooling layer with a pixel window size of 2×2 and a step size of 2 is utilized to reduce computational complexity efficiently. The convolutional layer's channel width doubles after each maximum pooling operation, starting with 32 channels and rising through the layers to reach 512 channels. The image size is halved using the ultimate pooling operation. The ViT architecture receives the output from low-level feature extraction. The (C, H, W) tensor can represent the feature extraction's internal state. A $512 \times 7 \times 7$ feature of the input image results from the feature extraction.

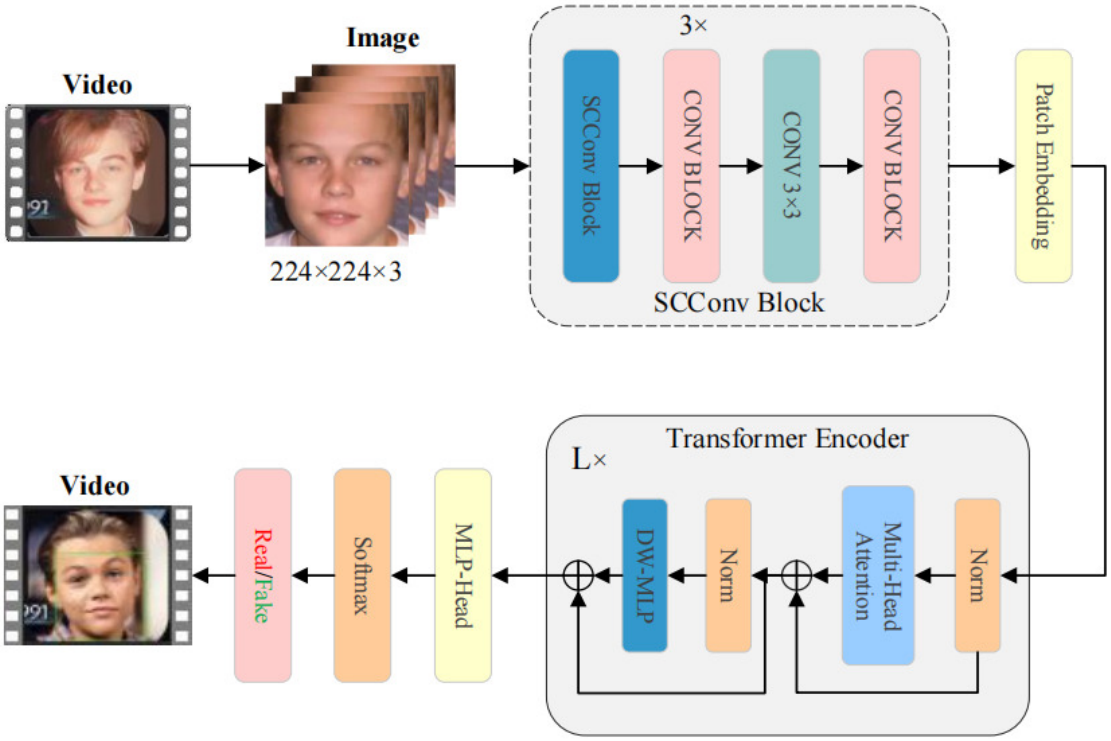

Figure 4. The process of detecting forged video using SCViT.

ViT applies transformer to a series of image blocks to further hone the picture categorization task. Following the creation of sequences, the segmented image chunks are sent to the transformer's multi-head self-attention layer for feature extraction. The size of the image after SCConv is 7×7. The feature map is separated into 7×7 inch blocks, which are then assembled to form sequences and sent to the transformer's multi-head self-attention layer for feature extraction. Six transformer encoders are

configured, and Cls Token is used for categorization. Specifically, Figure 4 depicts the entire network.

3.2.2. Transformer encoder improvement

Our encoder is unique from the original transformer encoder because it has DW-MLP [44] and multi-head self-attention [13] blocks. Zero-padded positional coding replaces fixed-size positional coding and adds a DW convolution with padding size one between the feed-forward network's first fully connected (FC) layer and the GELU. The multi-layer perceptron (MLP) block comprises a GELU activation layer, a fully connected layer, and a depthwise (DW) convolution.

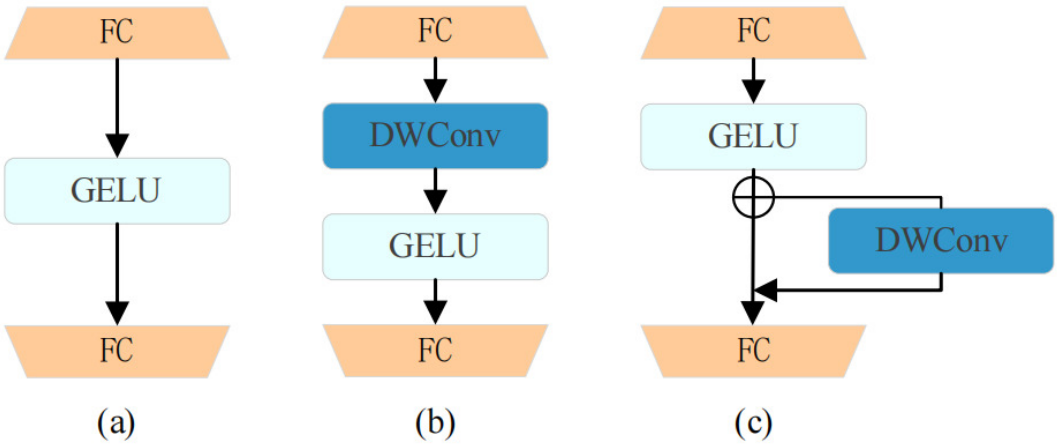

Figure 5. Improved structure of the MLP layer.

As shown in Figure 5, a is the original MLP layer, b and c are the improved MIP layer, and the enhanced model is used as our SCViTDW model. We also apply the DW convolution with added jump connections to the MLP layer, and the enhanced model is used as our SCViTDS model. The output layer consists of 2 channels, denoting the two classes associated with genuine and synthetic facial features. In contrast, the input layer encompasses 2048 channels. For the identification of the counterfeit class, the Softmax function operates on the output from the MLP header, producing a probability value within the range of 0 to 1. The proximity of the score to 1 serves as an indicator of enhanced model performance. The following formula, which employs the logarithmic loss function, is used:

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n}\Big[x_i \log(y_i) + (1 - x_i)\log(1 - y_i)\Big] \tag{3.1}$$

3.2.3. SCConv

We intend to utilize a CNN for the feature extraction component. However, conventional CNN architectures are characterized by substantial computational and storage requirements. The structured convolutional convolution (SCConv) method, as illustrated in Figure 6, is adapted to facilitate feature extraction efficacy while concurrently mitigating model parameters and FLOP counts. This enhancement is achieved by incorporating two distinct modules within the SCConv framework: the spatial redundancy

reduction unit (SRU) and the channel reconstruction unit (CRU). The SRU specifically addresses spatial redundancy, thus optimizing computational efficiency.

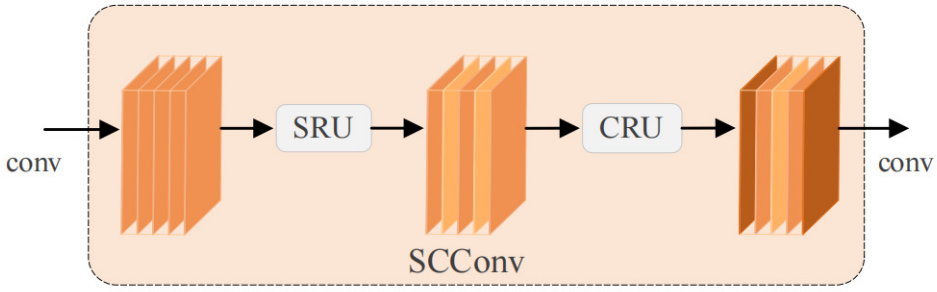

Figure 6. The SCConv architecture is seamlessly integrated with the SRU and CRU.

SRU uses separation-reconstruction operations. The group normalization (GN) layer's scaling factor is used in the separation operation to compare the information content of various feature maps. This allows us to analyze feature maps with various information contents separately. We initially normalize the input features $X \in \mathbb{R}^{N\times C\times H\times W}$ by dividing by the standard deviation σ and removing the mean μ. γ and β is the trainable affine transformation, with an added tiny positive constant ε for division stability, where N denotes the batch axis, C represents the channel axis, and H and W denote the axes of spatial height and width, respectively. The specifics are displayed below:

$$X_{out} = GN\left(X\right) = \gamma\frac{X-\mu}{\sqrt{\sigma^2+\varepsilon}} + \beta \tag{3.2}$$

The trainable parameters $\gamma \in \mathbb{R}^C$ within the GN layer are subsequently utilized to compute the variance of spatial pixels within each batch and channel. A higher resultant γ value signifies a more comprehensive representation of geographical information. The normalized correlation weights $W_\gamma \in \mathbb{R}^C$ are again applied to the feature map using a sigmoid function, which converts the weight values to ranges (0, 1). The specifics are displayed below:

$$W_\gamma = \left\{\omega_i\right\} = \frac{\gamma_i}{\sum_{j=1}^{C}\gamma_i}, i, j = 1, 2, \cdots, C \tag{3.3}$$

The process of gating is completed by thresholding for various weight values. Parts with less information will be suppressed and deemed redundant in this way. Through the amalgamation of informative features with less informative ones during the reconstruction operation, we achieve the generation of more information-rich features. Reducing redundant features in the spatial dimension contributes to refining representative features.

$$W = Gate\left(Sigmoid\left(W_\gamma\left(GN\left(X\right)\right)\right)\right) \tag{3.4}$$

Channel redundancy is handled using the CRU. CRU uses the split-transform-fuse operation. First, we convolutionally compress the αC channel and $(1-\alpha)C$ channel portions of the input feature X^w. After separating the upper X_{up} and lower X_{low} portions of the compressed features, X_{up} is then utilized in the top transformation step. This approach reduces computational costs by employing efficient

convolution techniques such as GWC and PWC instead of standard convolution. X_{low} as a complement to the enriched feature extractor is then passed into the bottom transformation stage, where the PWC operation is used to create feature maps with shallowly buried information. The final transformation stage might be written as follows:

$$Y_1 = M^G X_{up} + M^{P_1} X_{up} \tag{3.5}$$

The learnable weight matrix for GWC and PWC is $M^G \in \mathbb{R}^{\frac{\alpha c}{gr} \times k \times k \times c}$, $M^{P_1} \in \mathbb{R}^{\frac{\alpha c}{r} \times 1 \times 1 \times c}$ and the uppermost input and output feature maps are denoted as $X_{up} \in \mathbb{R}^{\frac{\alpha c}{r} \times h \times w}$ and $Y_1 \in \mathbb{R}^{c \times h \times w}$, respectively. The output of the following level is created by connecting the generated and reused features as follows:

$$Y_2 = M^{P_2} X_{low} \cup X_{low} \tag{3.6}$$

The learnable weight matrix of the PWC is denoted as $M^{P_2} \in \mathbb{R}^{\frac{(1-\alpha)c}{r} \times 1 \times 1 \times (\frac{1-\alpha}{r})c}$, with $X_{low} \in \mathbb{R}^{\frac{(1-\alpha)c}{r} \times h \times w}$ and $Y_2 \in \mathbb{R}^{c \times h \times w}$ representing the corresponding lower input and output feature maps. Then, channel statistics are used to capture global spatial information [45] using global average pooling, and the upper and lower features are combined using channels to produce channel-refined features. With channel statistics, global average pooling is used to gather global geographic information $S_m \in \mathbb{R}^{c \times 1 \times 1}$, which is calculated as:

$$S_m = Pooling(Y_m) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} Y_c(i, j), m = 1, 2 \tag{3.7}$$

The higher and lower global channel S_1, S_2 descriptions are up next are layered one on top of the other, and the feature importance vector $\beta_1, \beta_2 \in \mathbb{R}^c$ is produced using the channel soft attention procedure as follows:

$$\beta_1 = \frac{e^{S_1}}{e^{S_1} + e^{S_2}}, \beta_2 = \frac{e^{S_2}}{e^{S_1} + e^{S_2}}, \beta_1 + \beta_2 = 1 \tag{3.8}$$

The upper and lower features Y_1, Y_2 can then be combined channel by channel, guided by the feature importance vector β_1, β_2, to produce channel refinement features:

$$Y = \beta_1 Y_1 + \beta_2 Y_2 \tag{3.9}$$

As a component of our backbone network, SCConv comprises sequential connections of SRUs and CRUs, which can enhance model detection performance while lowering computation and storage.

4. Experiments

4.1. Dataset

As shown in Table 2, we acquired 89,635 face images from the Celeb-DF dataset, 179,016 images from the DFDC dataset, and 99,906 images from the FaceForensics++ dataset. The FaceForensics++ dataset includes versions with different compression rates: raw quality (quantization = 0), high quality (HQ, quantization = 23), and low quality (LQ, quantization = 40). We chose to train and test the model on FF++ (LQ). We chose these three datasets as our base dataset. The face images are stored in JPG files with consistent image quality. After scaling the dataset, we partitioned it into training, validation, and testing sets. The Celeb-DF and FaceForensics++ datasets were divided in a ratio of 70:15:15. In contrast, the DFDC dataset was divided in a ratio of 70:20:10. A substantial number of datasets, ample for training the model, have been amassed.

Table 2. Outcome of dataset partitioning.

Datasets	Real	Fake	Train	Validation	Test
Celeb-DF	44,819	44,816	62,747	13,445	13,443
FaceForensics++	49,962	49,944	69,935	14,986	14,985
DFDC	90,024	88,992	120,296	41,928	16,792

4.2. Experiment setting

The experimental hardware comprised an Intel (R) Xeon (R) CPU E5-2630 v3 @2.40GHz and NVIDIA GeForce RTX 3090 GPU, with debugging conducted using the PyTorch framework. Before entering the network, ensuring that the input image has been scaled to 224×224 pixels is imperative. Training occurs over 50 epochs, employing a batch size of 16. The optimization utilizes an Adam optimizer with a learning rate of 0.1×10^{-3} and a weight decay rate of 0.1×10^{-6}.

4.3. Evaluation metrics

1) Accuracy

A measurement used to assess classification models is accuracy. It is merely the proportion of all the model's accurate predictions produced. It can be determined using positive and negative categories in binary categorization in the manner shown below:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{4.1}$$

In the context of the presented metrics, FP represents false positive cases, TN denotes true negative cases, FP signifies false positive cases, and FN corresponds to false negative cases.

2) AUC

The receiver operating characteristic (ROC) curve illustrates the relationship between true positive rate (TPR) and false positive rate (FPR), with the Area Under the Curve (AUC) quantifying this correlation. FPR is plotted along the ROC curve's horizontal axis, while TPR is represented on the vertical axis. Mathematically, their relationship is expressed as:

$$FPR = \frac{FP}{FP + TN} \tag{4.2}$$

$$TPR = \frac{TP}{TP + FN} \tag{4.3}$$

3) F1-score

A statistic called the F1Score is used in statistics to assess a model's accuracy in binary classification, also known as multi-task binary classification. It considers the categorization model's recall as well as accuracy. The F1 Score constitutes a weighted average encompassing the model's precision and recall metrics, encompassing values from 0 to 1, where higher values correspond to superior model performance. The F1 Score is calculated using the following formula:

$$P = \frac{TP}{TP + FP} \tag{4.4}$$

$$R = \frac{TP}{TP + FN} \tag{4.5}$$

$$F1 - score = 2 \times \frac{P \times R}{P + R} \tag{4.6}$$

4) Params

The parameter count indicates the total number of parameters that undergo training during the model training process. This metric serves as a measure of the model's size, reflecting its computing space complexity.

5) Floating point operations

When calculating FLOPS, we typically add, subtract, multiply, divide, find the power, find the square root, and do other operations as a single FLOP to count. However, the number of floating point operations, understood as the amount of computation (computational time complexity), can be used to measure the algorithm's complexity.

4.4. Experimental results

The efficacy of the suggested SCViT network is assessed in this section. We train different models on the same dataset using the same training procedure, settings, and assessment metrics.

4.4.1. Comparing other approaches

Table 3 summarizes the follow-up tests performed on the FaceForensics++ dataset with 7490 real and manipulated facial images. The SCViTDW model shows impressive performance with an accuracy of 99.23%, an AUC of 0.9995, and an F1 score of 99.23%. Meanwhile, the SCViTDS model also performed very well, with an accuracy of 99.21%, an AUC of 0.9996, and an F1 score of 99.21%. The effectiveness of the enhanced model is highlighted by a comparative analysis with previous methods, especially the SCViT architecture that combines the SCConv convolution and ViT components.

Table 3. Comparison of different model results on the FaceForensics++ dataset.

Method	Acc	AUC	F1-score
Xception	94.95	0.9911	94.87
EfficientNetB0	78.83	0.8723	78.07
EfficientNetV2S	87.84	0.9515	88.08
MesoInception	65.06	0.7003	64.22
MesoNet	56.78	0.6347	35.00
EfficientNetB0ViT	85.62	0.9361	85.07
SCViT(ours)	99.07	0.9994	99.07
SCViTDS(ours)	99.21	0.9996	99.21
SCViTDW(ours)	99.23	0.9995	99.23

In addition, the SCViTDW and SCViTDS implemented by combining deep convolution (DW) perform excellently in accuracy, further confirming the model's effectiveness. The findings on the DFDC dataset are shown in Table 4, which includes 17,355 real and fake facial images. Of particular note, SCViTDW performs well in all metrics with an accuracy of 87.92%, an AUC of 0.9533, and an F1

score of 88.11%, second only to the highest score. Table 5 demonstrates the performance comparison when using the normal convolutional layer cascade ViT and adding DS convolution, and after adding SC convolution. The results show that SCViTDW and SCViTDS improve the test accuracy while reducing the computational effort, with FLOPs reduced from 6221.34 M to 5855.26 M and Params reduced from 88.52 to 88.51.

Table 4. Comparison of different model results on the DFDC dataset.

Method	Acc	AUC	F1-score
Xception	90.14	0.9754	90.16
EfficientNetB0	80.66	0.9107	80.57
EfficientNetV2S	85.03	0.9379	84.57
MesoInception	60.71	0.7543	44.18
MesoNet	61.68	0.7037	58.14
EfficientNetB0ViT	83.61	0.9206	83.05
SCViT(ours)	86.12	0.9425	86.41
SCViTDS(ours)	87.76	0.9537	87.98
SCViTDW(ours)	87.92	0.9533	88.11

Table 5. Experimental results of different convolutional modules cascaded with ViT on the DFDC dataset.

	module			DFDC				
Method	SC	DS	DW	Acc	AUC	F1-score	FLOPs (M)	Params (M)
SC + ViT	✓			86.12	0.9425	86.41	5855.04	88.39
ViT + DS		✓		87.97	0.9518	88.10	6221.34	88.52
SC + ViT + DS	✓	✓		87.76	0.9537	87.98	5855.26	88.51
SC + ViT + DW	✓		✓	87.92	0.9533	88.11	5855.26	88.51

Table 6. Experimental results of different convolutional modules cascaded with ViT on the Celeb-DF dataset.

	module			Celeb-DF		
Method	SC	DS	DW	Acc	AUC	F1-score
SC + ViT	✓			99.51	0.9954	99.51
SC + ViT + DS	✓	✓		100.00	1.000	100.00
SC + ViT + DW	✓		✓	99.98	0.9998	99.98

Finally, we tested on the Celeb-DF dataset, and Table 6 showed impressive results. The accuracy of SCViTDW and SCViTDS are close to 100%, highlighting the remarkable performance of the models in terms of training accuracy and computational efficiency. We further evaluated our approach against other state-of-the-art models. As shown in Table 7, the best performance in terms of AUC metrics is achieved on the FF++ and Celeb-DF datasets. With the second-highest AUC value on the DFDC

dataset, the results demonstrate the practical generalization ability of our framework.As shown in Table 8, not only does our model outperform others in terms of AUC values, but it also boasts lower FLOPs and parameters. This further highlights the efficiency and superiority of our method. The results of other works are mainly cited from [32,33,38,46–53].

Table 7. Comparison with state-of-the-art methods on FF++, CelebDF and DFDC.

	Test Set AUC			
Method	DFDC	FF++	Celeb-DF	Avg
SBIs [32]	0.7242	0.9964	0.9318	0.8841
UCF [33]	0.8050	0.9960	0.8240	0.8750
M2TR [46]	-	0.9531	0.9980	0.9756
MARLIN [47]	-	0.9305	0.9561	0.9433
TALL-Swin [48]	0.7678	0.9457	0.9079	0.8738
Face X-ray [49]	0.6550	0.6160	0.7950	0.6887
LipForensics [4]	0.7350	0.9810	0.8240	0.8460
RealForensics [50]	0.7590	0.9950	0.8690	0.8743
Xception [28]	**0.9754**	0.9910	0.9027	0.9564
SCViTDW (ours)	0.9533	**0.9995**	**0.9998**	**0.9842**

Table 8. Performance of different models.

Method	DFDC	Celeb-DF	FLOPs (G)	Params (M)
VidTR [51]	0.7330	0.8330	117.0	93
ISTVT [52]	0.7420	0.8410	455.8	-
VTN [53]	0.7350	0.8320	296.6	46
TALL-Swin [48]	0.7678	0.9079	47.5	86
EfficientNetB0ViT	0.9206	-	-	109
SCViTDW(ours)	**0.9533**	**0.9998**	5.9	89

4.4.2. Visualized results

We can observe the inferred outcomes for authentic and fake faces in Figure 7. Above the photographs are the anticipated categories and scores. The images in the first row are fake, while the pictures in the second row are real. The exemplary performance of a model is notably exemplified by its proximity to a score of 1. Consequently, our detection categories and scores closely resembling the actual scenario demonstrate our hybrid model's exceptional performance. As shown in Figures 8 and 9, The detection procedure concludes with synthesizing a novel video output after extracting facial features from the input video. Subsequently, the extracted faces undergo scrutiny through our deepfake detection model, wherein the detection outcomes are annotated per frame. Ultimately, our method achieves the discrimination between authentic and manipulated videos. The precision of our proposed approach is manifested through the meticulous delineation of the actual state of each face in the cinematic sequence, as discerned through a frame-by-frame presentation of the detection results. As illustrated in Figure 10,

to exemplify the comprehensiveness of our approach, we conducted a comparative analysis of three metrics inherent to our model against those of other state-of-the-art models for deepfake detection. This evaluation was performed on two distinct datasets. The outcomes of our investigation strongly indicate the efficacy of our method. Additionally, we present detailed ROC curves for each model, elucidating their performance characteristics in Figure 11. Our ROC curve is the second closest to the top left corner of the DFDC dataset, and it is most relative to the entire left curve on the FaceForensics++ dataset. This indicates that our model performs better than the competition.

Figure 7. The detection outcomes for facial images, where the first row encompasses counterfeit images, and the second row comprises authentic images.

Figure 8. Frame-by-frame detection outcomes for genuine videos.

4.4.3. Discussion

Recently, researchers have proposed a series of innovative approaches to address the limitations of deep forgery detection models regarding accuracy and localization. For example, in [37], they offered

Figure 9. Frame-by-frame detection outcomes for fake videos.

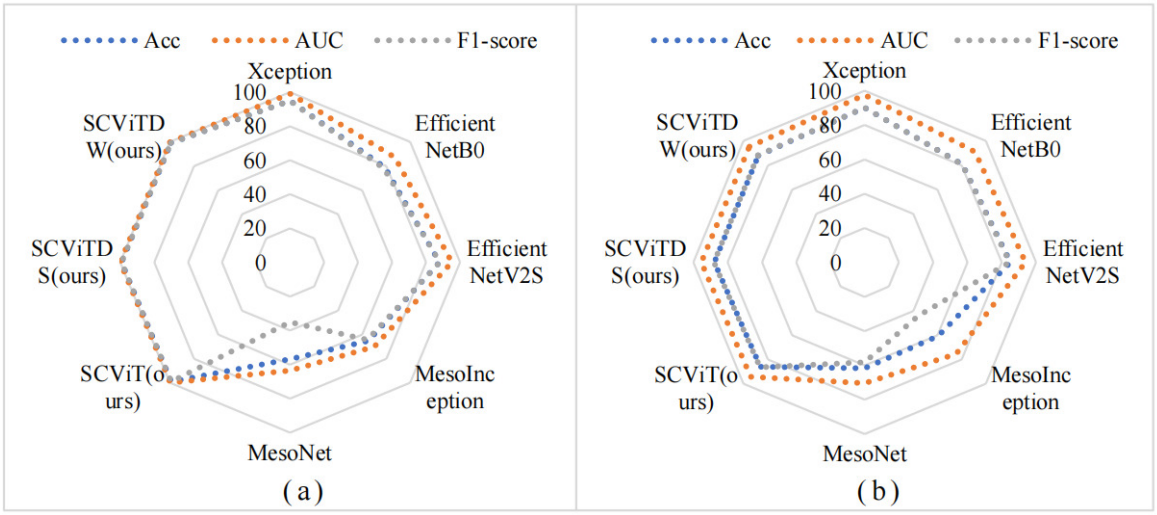

Figure 10. Performance comparison on different datasets. (a) shows three performance comparisons between our model and other models on FaceForensics++ dataset, and (b) shows three performance comparisons between our model and its model on DFDC dataset.

the detect any deepfakes (DADF) framework with multi-scale adapters based on SAM, efficiently fine-tuned by capturing both short-term and long-term forgery contexts. Although the model achieved an average accuracy of 95.94% on the FF++ dataset, our model achieves superior results on the same dataset, reaching an accuracy of 99.23%. Meanwhile, other scholars have proposed the Xception LSTM algorithm to capture and enhance spatio-temporal correlations before Xception downscaling. Although our method achieved better results on the Celeb-DF dataset than [54], it failed to outperform the results of [54] on the DFDC dataset. For the video-level detection of forged faces, [38] utilized a new contrast spatiotemporal extraction method to improve the detection of high compression depth generated videos through fine-grained spatial frequency cues and temporal contrasts, improving model stability.

Our model achieves better results in predicting Celeb-DF, FaceForensics++, and DFDC datasets and demonstrates the effectiveness of cascade networks in deep forgery detection, surpassing other cutting-edge deep forgery detection techniques. The results also validate the improvement of test frames

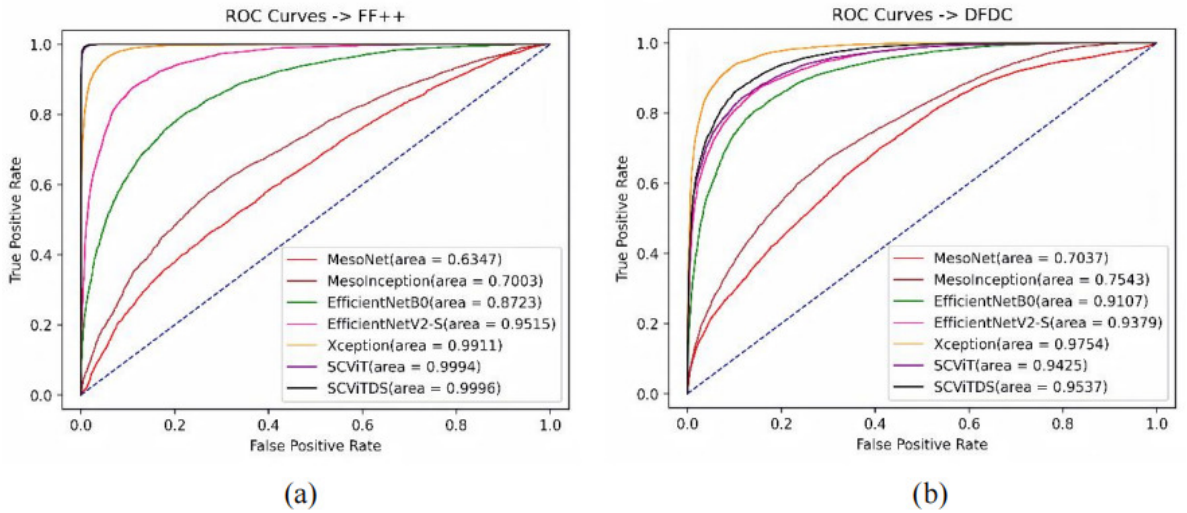

Figure 11. ROC curves comparing our model with others. (a) ROC curves on the FaceForensics++ dataset. (b) ROC curves on the DFDC dataset.

when using DW convolution, showing that DW plays an active role in our work. It is worth noting that when we jump-connect the DW convolution, we can achieve 100% accuracy on the Celeb-DF dataset, and the spatial and channel-based reconstruction convolution (SCConv) can effectively alleviate the computational burden of the model and reduce the computational amount of the model, demonstrating the broad applicability of our cascade network. In addition, the model excels in its ability to differentiate between still images and dynamic video sequences and addresses various classification challenges by utilizing alternative structural configurations in the feature extraction module of the detection model.

To better address future research directions, we propose the following recommendations: enhance research on 3D faces, including the creation of more 3D face datasets covering different populations, ages, and ethnicities to improve the generalization performance of the model; consider enhancing the robustness of models in dynamic environments so that they can effectively detect moving and changing 3D faces; promote the synthesis of multimodal data to improve the model's understanding of natural scenes and enhance the experience of new forgery method modes in response to continuously updated forgery techniques to build more effective detection systems.

5. Conclusions

This paper presents a deepfake detection model structured as a cascaded network, incorporating a ViT and an SCConv convolution module. Our dataset for fake video detection is curated through the application of the BlazeFace face detector to identify facial regions within the dataset. The ViT effectively assimilates global and local features, with the initial convolutional segment dedicated to feature extraction from the input image. The incorporation of SCConv serves to curtail the number of parameters and computational workload, thereby enhancing the efficiency of the feed-forward neural layer. Furthermore, it augments the adaptability of the transformer encoder to accommodate arbitrary input sizes. We have added zero-padded positional coding to the ViT feed-forward network to improve the model's performance. On the FaceForensics++ dataset, our model attained an AUC of

0.9995, an accuracy (Acc) of 99.23%, and an F1 score of 99.23%. Likewise, on the DFDC dataset, our model exhibited an AUC of 0.9533, an accuracy of 87.92%, and an F1 score of 88.11%. On the Celeb-DF dataset, our model shows AUC, accuracy, and F1 scores close to 100 percent. Furthermore, enhancements were made to the feed-forward neural layer to augment the adaptability of the transformer encoder in handling diverse sizes, consequently reducing both the number of model computations and parameters. Lastly, to accurately verify the authenticity of deepfake films, we also carried out visual inspection at the picture and video levels. We plan to enhance the accuracy of our detection model in further work.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

Harbin Normal University Postgraduate Innovative Research Project by Grant Number HSDSSCX2022-144.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. V. Kumar, V. Kansal, M. Gaur, Multiple forgery detection in video using convolution neural network, *Comput. Mater. Continua*, **73** (2022), 1347–1364. https://doi.org/10.32604/cmc.2022.023545
2. F. Ding, B. Fan, Z. Shen, K. Yu, G. Srivastava, K. Dev, et al., Securing facial bioinformation by eliminating adversarial perturbations, *IEEE Trans. Ind. Inf.*, **19** (2023), 6682–6691. https://doi.org/10.1109/TII.2022.3201572
3. A. Ilderton, Coherent quantum enhancement of pair production in the null domain, *Phys. Rev. D*, **101** (2020), 016006. https://doi.org/10.1103/physrevd.101.016006
4. A. Ilderton, Lips don't lie: A generalisable and robust approach to face forgery detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 5039–5049. https://doi.org/10.1109/CVPR46437.2021.00500
5. N. Yu, L. Davis, M. Fritz, Attributing fake images to gans: Learning and analyzing gan fingerprints, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 7556–7566. http://doi.org/10.1109/ICCV.2019.00765
6. N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, S. Tubaro, Video face manipulation detection through ensemble of CNNs, in *2020 25th International Conference on Pattern Recognition (ICPR)*, (2021), 5012–5019. http://doi.org/10.1109/ICPR48806.2021.9412711
7. H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, N. Yu, Multi-attentional deepfake detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 2185–2194. http://doi.org/10.1109/CVPR46437.2021.00222

8. J. Li, Y. Wen, L. He, SCConv: Spatial and channel reconstruction convolution for feature redundancy, in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2023), 6153–6162. http://doi.org/10.1109/CVPR52729.2023.00596

9. J. Kim, J. K. Lee, K. M. Lee, Accurate image super-resolution using very deep convolutional networks, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 1646–1654. http://doi.org/10.1109/CVPR.2016.182

10. E. Zakharov, A. Shysheya, E. Burkov, V. Lempitsky, Few-shot adversarial learning of realistic neural talking head models, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 9458–9467. http://doi.org/10.1109/ICCV.2019.00955

11. R. Haridas, L. Jyothi, Convolutional neural networks: A comprehensive survey, *Int. J. Appl. Eng. Res.*, **14** (2019), 780. http://doi.org/10.37622/IJAER/14.3.2019.780-789

12. K. R. Prajwal, R. Mukhopadhyay, P. J. Philip, A. Jha, V. Namboodiri, C. V. Jawahar, Towards automatic face-to-face translation, in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019. http://doi.org/10.1145/3343031.3351066

13. K. R. Prajwal, R. Mukhopadhyay, V. Namboodiri, C. V. Jawahar, A lip sync expert is all you need for speech to lip generation in the wild, in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. http://doi.org/10.1145/3394171.3413532

14. Y. Nirkin, L. Wolf, Y. Keller, T. Hassner, DeepFake detection based on discrepancies between faces and their context, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2022), 6111–6121. http://doi.org/10.1109/TPAMI.2021.3093446

15. Z. Xu, Z. Hong, C. Ding, Z. Zhu, J. Han, J. Liu, et al., Mobilefaceswap: A lightweight framework for video face swappingg, preprint, arXiv:2201.03808. https://doi.org/10.48550/arXiv.2005.07034

16. T. Wang, Z. Li, R. Liu, Y. Wang, L. Nie, An efficient attribute-preserving framework for face swapping, *IEEE Trans. Multimedia*, **44** (2024), 1–13. http://doi.org/10.1109/TMM.2024.3354573

17. B. Peng, H. Fan, W. Wang, J. Dong, S. Lyu, A unified framework for high fidelity face swap and expression reenactment, *IEEE Trans. Circuits Syst. Video Technol.*, **32** (2022), 3673–3684. http://doi.org/10.1109/TCSVT.2021.3106047

18. H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, Z. Liu, Pose-controllable talking face generation by implicitly modularized audio-visual representation, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 4174–4184. http://doi.org/10.1109/CVPR46437.2021.00416

19. N. Van Huynh, D. T. Hoang, D. N. Nguyen, E. Dutkiewicz, DeepFake: Deep dueling-based deception strategy to defeat reactive jammers, *IEEE Trans. Wireless Commun.*, **20** (2021), 6898–6914. https://doi.org/10.1109/TWC.2021.3078439

20. A. Hamza, A .R. R. Javed, F. Iqbal, N. Kryvinska, A. S. Almadhor, Z. Jalil, et al., Deepfake audio detection via MFCC features using machine learning, *IEEE Access*, **10** (2022), 134018–134028. http://doi.org/10.1109/ACCESS.2022.3231480

21. S. Bounareli, C. Tzelepis, V. Argyriou, I. Patras, G. Tzimiropoulos, HyperReenact: one-shot reenactment via jointly learning to refine and retarget faces, in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, **10** (2023), 7115–7125. http://doi.org/10.1109/ICCV51070.2023.00657

22. F. T. Hong, L. Shen, D. Xu, Depth-aware generative adversarial network for talking head video generation, *IEEE Trans. Pattern Anal. Mach. Intell.*, **10** (2023), 1–15. http://doi.org/10.1109/TPAMI.2023.3339964

23. N. Liu, F. Zhang, L. Chang, F. Duan, Scattering-based hybrid network for facial attribute classification, *Front. Comput. Sci.*, **10** (2024). http://doi.org/10.1007/s11704-023-2570-6

24. Y. Xu, Y. Yin, L. Jiang, Q. Wu, C. Zheng, C. C. Loy, et al., Transeditor: Transformer-based dual-space gan for highly controllable facial editing, preprint, arXiv:2203.17266. https://doi.org/10.48550/arXiv.2203.17266

25. J. Sun, X. Wang, Y. Zhang, X. Li, Q. Zhang, Y. Liu, et al., Fenerf: Face editing in neural radiance fields, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 7662–7672. http://doi.org/10.1109/CVPR52688.2022.00752

26. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 770–778. http://doi.org/10.1109/CVPR.2016.90

27. G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 2261–2269. http://doi.org/10.1109/CVPR.2017.243

28. F. Chollet, Xception: Deep learning with depthwise separable convolutions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 1800–1807. http://doi.org/10.1109/CVPR.2017.195

29. M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in *International Conference on Machine Learning*, PMLR, (2019), 6105–6114.

30. D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, Mesonet: a compact facial video forgery detection network, in *2018 IEEE Iinternational Workshop on Information Forensics and Security (WIFS)*, 2018. http://doi.org/10.1109/wifs.2018.8630761

31. T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, W. Xia, Learning self-consistency for deepfake detection, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 15003–15013. http://doi.org/10.1109/ICCV48922.2021.01475

32. K. Shiohara, T. Yamasaki, Detecting deepfakes with self-blended images, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 18699–18708. http://doi.org/10.1109/CVPR52688.2022.01816

33. Z. Yan, Y. Zhang, Y. Fan, B. Wu, UCF: Uncovering common features for generalizable deepfake detection, preprint, arXiv:2304.13949. https://doi.org/10.48550/arXiv.2304.13949

34. Y. Xu, K. Raja, L. Verdoliva, M. Pedersen, Learning pairwise interaction for generalizable deepFake detection, preprint, arXiv:2302.13288. https://doi.org/10.48550/arXiv.2302.13288

35. B. Huang, Z. Wang, J. Yang, J. Ai, Q. Zou, Q. Wang, et al., Implicit identity driven deepfake face swapping detection, in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2023), 4490–4499. https://doi.org/10.1109/CVPR52729.2023.00436

36. Y. Lai, Z. Luo, Z. Yu, Detect any deepfakes: Segment anything meets face forgery detection and localization, preprint, arXiv:2306.17075. https://doi.org/10.48550/arXiv.2306.17075

37. Y. Zhu, C. Zhang, J.Gao, X. Sun, Z. Rui, X. Zhou, High-compressed deepfake video detection with contrastive spatiotemporal distillation, *Neurocomputing*, **565** (2024),126872. https://doi.org/10.1016/j.neucom.2023.126872
38. L. Deng, J. Wang, Z. Liu, Cascaded network based on efficientNet and transformer for deepfake video detection, *Neural Process. Lett.*, **55** (2023). http://doi.org/10.1007/s11063-023-11249-6
39. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Niessner, Faceforensics++: Learning to detect manipulated facial images, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 1–11. http://doi.org/10.1109/ICCV.2019.00009
40. B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, et al., The deepfake detection challenge (DFDC) dataset, preprint, arXiv:2006.07397. https://doi.org/10.48550/arXiv.2006.07397
41. Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-df: A large-scale challenging dataset for deepfake forensics, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 3204–3213. http://doi.org/10.1109/CVPR42600.2020.00327
42. V. Bazarevsky, Y. Kartynnik, A. Vakunov, K. Raveendran, M.Grundmann, Blazeface: Sub-millisecond neural face detection on mobile GPUs, preprint, arXiv:1907.05047. https://doi.org/10.48550/arXiv.1907.05047
43. M. Diganta, Mish: A self regularized non-monotonic activation function, preprint, arXiv:1908.08681. https://doi.org/10.48550/arXiv.1908.08681
44. W. Wang, E. Xie, X. Li, D. Fan, K. Song, D. Liang, et al., Pvt v2: Improved baselines with pyramid vision transformer, *Comput. Visual Media*, **8** (2022), 415–424. https://doi.org/10.1007/s41095-022-0274-8
45. R. Congalton, Accuracy assessment and validation of remotely sensed and other spatial information, *Int. J. Wildland Fire*, **10** (2001), 321–328. http://doi.org/10.1071/WF01031
46. J. Wang, Z. Wu, W. Ouyang, X, Han, J, Chen, S. Lim, et al., M2tr: Multi-modal multi-scale transformers for deepfake detection, preprint, arXiv:2104.09770. https://doi.org/10.48550/arXiv.2104.09770
47. Z. Cai, S. Ghosh, K. Stefanov, A. Dhall, J. Cai, H. Rezatofighi, et al., Marlin: Masked autoencoder for facial video representation learning, preprint, arXiv:2211.06627. https://doi.org/10.48550/arXiv.2211.06627
48. Y. Xu, J. Liang, G. Jia, Z. Yang, Y. Zhang, R. He, TALL: Thumbnail layout for deepfake video detection, preprint, arXiv:2307.07494. https://doi.org/10.48550/arXiv.2307.07494
49. L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, et al., Face X-Ray for more general face forgery detection, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 5000–5009. https://doi.org/10.1109/CVPR42600.2020.00505
50. A. Haliassos, R. Mira, S. Petridis, M. Pantic, Leveraging real talking faces via self-supervision for robust forgery detection, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 14930–14942. https://doi.org/10.1109/CVPR52688.2022.01453
51. Y. Zhang, X. Li, C. Liu, B. Shuai, Y. Zhu, B. Brattoli, et al., Vidtr: Video transformer without convolutions, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2020), 13557–13567. https://doi.org/10.1109/ICCV48922.2021.01332

52. C. Zhao, C. Wang, G. Hu, H. Chen, C. Liu, J. Tang, ISTVT: interpretable spatial-temporal video transformer for deepfake detection, *IEEE Trans. Inf. Forensics Secur.*, (2023), 1335–1348. https://doi.org/10.1109/TIFS.2023.3239223

53. D. Neimark, O. Bar, M. Zohar, D. Asselmann, Video transformer network, in *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, (2023), 3156–3165. https://doi.org/10.1109/ICCVW54120.2021.00355

54. B. Chen, T. Li, W. Ding, Detecting deepfake videos based on spatiotemporal attention and convolutional LSTM, *Inf. Sci.*, **601** (2022), 58–70. https://doi.org/10.1016/j.ins.2022.04.014

AIMS Press