Received: 23 February 2021 | Revised: 27 April 2021 | Accepted: 11 May 2021

DOI: 10.1002/int.22499

WILEY

RESEARCH ARTICLE

# A lightweight 3D convolutional neural network for deepfake detection

Jiarui Liu[1] | Kaiman Zhu[1] | Wei Lu[1] | Xiangyang Luo[2] | Xianfeng Zhao[3,4]

[1]School of Computer Science and Engineering, Guangdong Province Key Laboratory of Information Security Technology, Ministry of Education Key Laboratory of Machine Intelligence and Advanced Computing, Sun Yat-sen University, Guangzhou, China

[2]State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, China

[3]State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China

[4]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

**Correspondence**
Wei Lu, School of Computer Science and Engineering, Sun Yat-sen University, 510006 Guangzhou, China.
Email: luwei3@mail.sysu.edu.cn

**Abstract**

The rapid development of DeepFake technologies has brought great challenges to the authenticity of video contents. It is of vital importance to develop DeepFake detection methods, among which three-dimensional (3D) convolution neural networks (CNN) have attracted wide interest and achieved satisfying performances. However, there are few 3D CNNs designed for DeepFake detection and the parameters of them are large, which cause heavy memory and storage consumption. In this paper, a lightweight 3D CNN is proposed for DeepFake detection. Channel transformation module is designed to extract features with much fewer parameters in higher level. Serving as spatial-temporal module, 3D CNNs are adopted to fuse the spatial features in time dimension. To suppress frame content and highlight frame texture, spatial rich model features are extracted from the input frames, which helps the spatial-temporal module achieve better performance. Experimental results show that the number of parameters of the proposed network is much less than those of other networks and the proposed network outperforms other state-of-the-art DeepFake detection methods on mainstream DeepFake data sets.

KEYWORDS

3D CNN, deepfake, deepfake detection, face swapping, face manipulation

# 1 | INTRODUCTION

With the development of video generation technique,[1,2] it is more convenient for the public to create realistic synthesized videos or modify content of videos nowadays, especially with the help of Generative Adversarial Networks (GAN).[3,4] Such video generation technique has been used for positive purposes like privacy preservation,[5-10] but could also cause social problems. In 2017, a Reddit user named "DeepFakes" used deep-learning-based DeepFake methods to create some pornographics videos with swapped faces of celebrities and published them online. Such an event has not only greatly injured the right of portrait and reputation of the involved people, but also taken it's toll on online security. In addition to fake pornographics, there are more harmful usages such as fake news, hoaxes and financial fraud. Moreover, if the faces of politicians are swapped into malicious talking videos, it could lead to conflicts among countries. Therefore, it is important to develop DeepFake detection methods to help judge the authenticity of the videos.

"DeepFake," which refers to a wild range of face manipulation techniques, includes the state-of-the-art methods such as computer vision[11,12] and deep learning.[13] In general, face manipulation can be categorized into four groups[14]: entire face synthesis, identity swap, attribute manipulation, and expression swap. Identity swap, also named as face swap,[11] is one of the most popular kinds of DeepFake videos, by which the faces of the source individuals are swapped on to the target individuals.

For security concerns, research are making efforts to expose DeepFake videos. Several DeepFake data sets have been published to promote the development of detection methods, such as DeepFake-TIMIT[15] and FaceForensics++.[16] Some detection methods take single frame as input sample, which are called frame level detection. Such methods extract features in spatial dimension such as the abnormal structure of the fake faces or the artifacts caused by blending fake faces into the frame. Rossler et al.[16] apply X-ception Net[17] initialized with the ImageNet weights and fine-tuned framework designed by Cozzolino et al.[18] which casts the hand-crafted steganalysis features to a convolution neural networks (CNN)-based network to verify the authenticity of faces cropped from frames. Afchar et al.[19] propose two different networks to expose fake videos focusing on their mesoscopic properties and then average the network prediction over the video. Li et al.[20] capture the distinctive artifacts caused by limited resolution of generated pictures and further warping step through CNNs to distinguished fake and real frame faces. Most of frame level methods treat each frame equally to produce result for videos, ignoring connections and differences among them.

To improve the accuracy and reliability on DeepFake video detection, some methods take consecutive frames as input samples, which are called video level detection. Such methods extract joint features in spatial and time dimensions. Sabir et al.[21] and Guera and Delp[22] propose detection methods with CNNs to extract frame level features followed by recurrent neural networks (RNN) to further learn the frame sequences. Ganiyusufoglu et al.[23] apply three-dimensional (3D) CNNs to consecutive frames and compare it with RNN-based methods, empirically showing greater performance. Injection of 3D CNNs into architectures seems to be a new trend in DeepFake detection. However, while 3D CNNs show promising performance on DeepFake detection, whether the amount of paraments they use is suitable is worthy of consideration. As many of 3D CNNs have to deal with a number of frames at the same time, the computation and the number of parameters are inevitably large, which leads to heavy memory and storage consumption.

In this paper, a lightweight 3D CNN is proposed. This model takes advantage of the excellent learning ability of the 3D CNNs in fusing spatial features in time dimension and uses a channel transformation (CT) module to make the number of parameters as small as possible while learning deeper level of the extracted features. Spatial rich model (SRM) features are also adopted to expose textures of the frames to further improve the performance of the spatial-temporal module. The result of experiments shows the proposed network reaches satisfying performance compared with other state-of-the-art DeepFake detection methods.

The contributions of this paper are summarized as follows:

- A lightweight 3D CNN is proposed for DeepFake detection. The lightweight network takes advantage of the excellent performance of the 3D CNNs on fusing spatial features in time dimensions.
- CT module is designed to extract features in higher level. CT and pointwise convolution are adopted to reduce the parameters of the network.
- The state-of-the-art performances on mainstream DeepFake detection data sets are achieved, which shows that the proposed network can efficiently detect DeepFake videos with much fewer parameters.

The paper is organized as follows: in Section 2, a glimpse of some current DeepFake detection methods both on frame level and video level and the discussion of the application of 3D CNNs are given; the proposed method is introduced in Section 3; experiments and analysis are discussed in Section 4; finally the conclusions are given in Section 5.

## 2 | RELATED WORKS

In this section, a brief summary of existing DeepFake detection methods is given in Section 2.1. And then 3D CNNs are introduced in Section 2.2.

### 2.1 | DeepFake detection

Detection on frame level means that frames of DeepFake videos are picked out as separated samples, and then intra-frame feature is learned. Such detection methods which focus on local spatial feature have potential to indicate manipulation face area.[24] Some works implicitly extract image features using CNN-based models[25,19,26] to directly make prediction. Noise of frame image is also found helpful for detection.[27] Visual artifacts of faces are also explored, including unnatural iris and teeth.[28] To give a better explanation for the prediction, multitask learning is found effective. Injecting manipulation masks reconstruction task into the networks[29] not only helps detection but also improves the generalization ability.[30] Reconstruction of input facial image is also an available assistant task.[31] Further more, zero-shot learning is a new direction to solve the unseen DeepFake types[32] as well. Methods mentioned above have explored spatial distinct features to differentiate fake and real face frames.

It is well recognized that temporal information plays an important role in video level detection.[33] Different temporal networks are proposed to overcome the shortages in existing versatile models and make them more suitable for DeepFake detection. For example, Xiao et al.[34] propose a dual-stage attention based ConvLSTM network for multivariate time series prediction. DeepFake detection on video level refers that prediction is made on videos (a number of consecutive frames can also be viewed as a small video) with information flowing between each frame and a joint representaion of frames is learned by the models. Considering sequential information in frames, RNN-based methods are put forward.[21,22] CNNs extract distinctive features of each frame at first, followed by some RNNs to capture inconsistencies between frames that flow over time. Chen et al.[33] also use CNNs as spatial feature extractor and RNNs as temporal feature aggregator, but learn the reconstruction of masks and verify the authenticity of each frame as well besides video prediction. While sequential models consist of two separable steps, some temporal models with 3D CNNs as backbone networks mix them together.

## 2.2 | 3D CNNs

3D CNNs[35] have been first proposed for action recognition. Its central idea of learning frames together within a certain period then is found useful in other video-based tasks. Some famous architectures such as C3D,[36] I3D,[37] R3D[38] and more sophisticated ones have been tested in DeepFake detection and show surprising performance.[23,39,40] In these works, detection of cross-manipulation methods is also a preferred task as 3D CNNs have been proved to be more generalized than RNNs based models, in addition to better performance on testing data set from the same source as in training. Choi et al.[41] further adapt predictive uncertainty and introduce attention mechanism to handle frame features before 3D CNNs. Although 3D CNNs seem to show good performance as they can learn the joint characteristics of in spatial and time dimensions, the paraments of them may beyond the requirements of DeepFake detection and thus easily overfit the training data set with heavy memory and storage consumption.

In the following section, we are going to show our proposed architecture which makes use of the advantage of 3D CNNs, but also reduces the parameters while maintaining comparable or even better performance than some other 3D CNNs on DeepFake detection.

# 3 | PROPOSED METHOD

In this section, the proposed lightweight 3D CNN is introduced. As shown in Figure 1, four consecutive frames are taken as input and then SRM features are extracted. The proposed network mainly consists of two modules including CT module and spatial-temporal module. First, the CT module is introduced in Section 3.1, which learns higher level of the extracted features. Then the spatial-temporal module is discussed in Section 3.2 and SRM features are introduced in Section 3.3. Finally, the overall framework is given in Section 3.4.

## 3.1 | CT module

The CT module aims to extract deep level features from the output of the spatial-temporal module with parameters as few as possible. As mentioned above, compared with traditional two-dimensional (2D) CNNs, 3D CNNs have larger parameters, which has a negative impact on

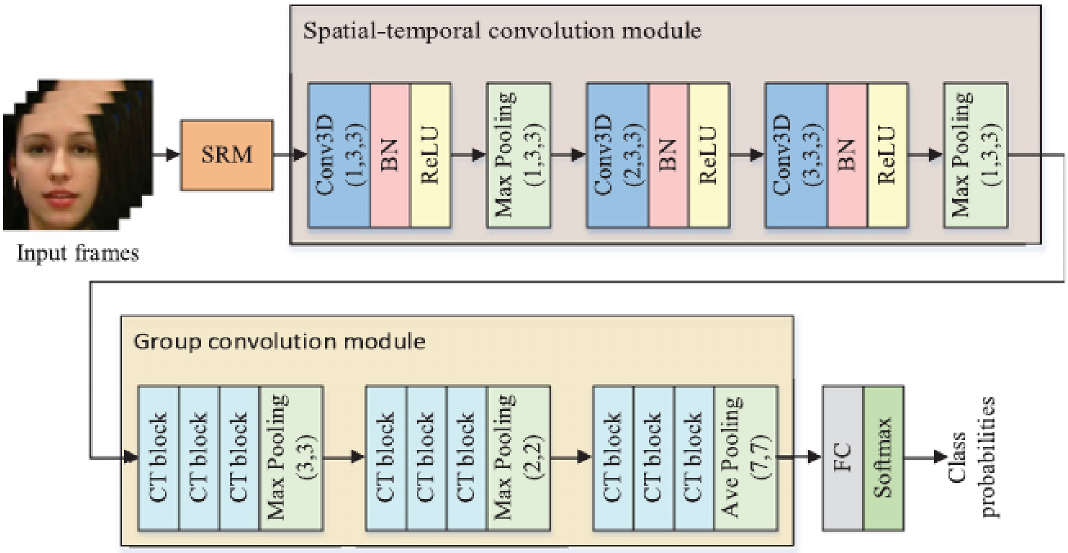

**FIGURE 1** Overview of the overall framework. The proposal takes four frames as input and extracts SRM features. Then the features are input into the proposed network and judge the authenticity of the faces. The proposed network mainly consists two modules including spatial-temporal module and CT module. CT, channel transformation; SRM, spatial rich model [Color figure can be viewed at wileyonlinelibrary.com]

the convergence and generalization capabilities. To reduce the number of parameters, the CT module is designed to replace the normal 3D convolution layers.

The module consists of nine CT blocks to extract deep level features, whose structure is shown in Figure 2. The left-side of the red arrow is the structure of the CT block and the right-side is the schematic diagram. The first pointwise convolutional layer with $1 \times 1$ kernels is mainly used to reduce channels to 1/4 of the input channels. Then $3 \times 3$ kernels are adopted to perform feature extraction using depthwise separable convolution[17] on each channel separately. With the help of pointwise convolution,[17] the channels of the feature maps are reflected to the output channels. The feature maps are finally processed with batch normalization layer and ReLU layer.

The CT block reduces the number of parameters by reducing the number of channels. What's more, depthwise separable convolution performs convolution on each channel with the same convolution kernel, which also reduce the parameters significantly. The number of parameters $N$ of a convolutional layer is defined as follows:

$$N = c_{in} \times k \times k \times c_{out} \tag{1}$$

where $c_{in}$ is the number of input channels and $k$ is the size of the convolutional kernel and $c_{out}$ is the number of output channels. The number of parameters of the three convolutional layers $N_1$, $N_2$, and $N_3$ are listed as follows:

$$N_1 = c_{in} \times 1 \times 1 \times c_{in}/4 \tag{2}$$

$$N_2 = c_{in}/4 \times 3 \times 3 \tag{3}$$

$$N_3 = c_{in}/4 \times 1 \times 1 \times c_{out} \tag{4}$$

In depthwise separable convolutional layer, one convolution kernel is responsible for one channel, and each channel is convoluted by only one corresponding convolution kernel. As a

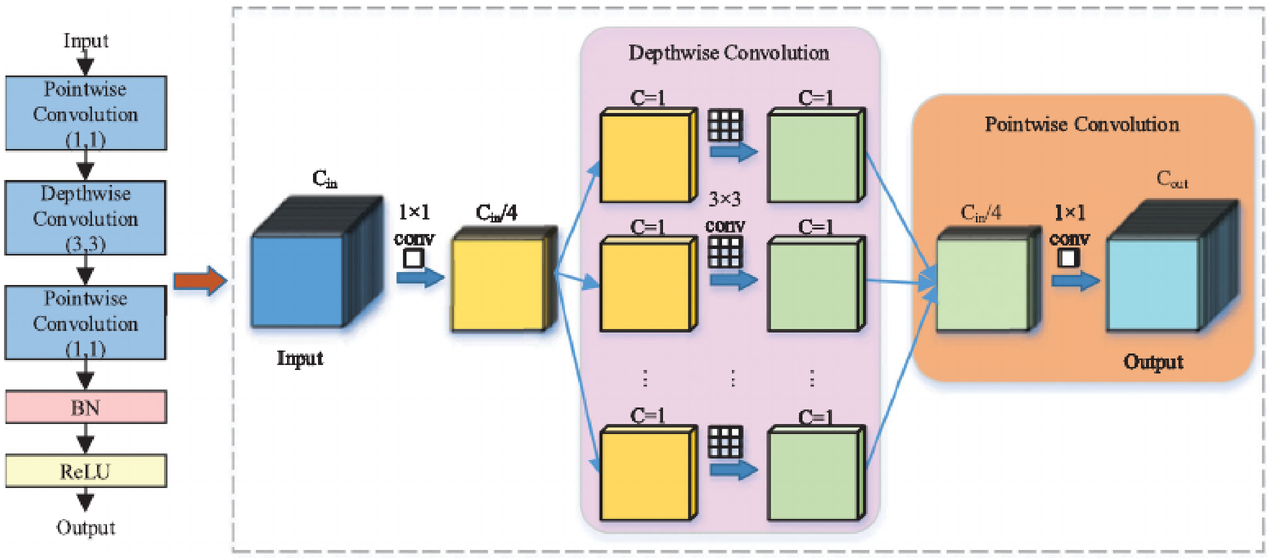

**FIGURE 2** Overview of the CT block. The left side of the red arrow is the structure of the CT block and the right side is the schematic diagram of the CT block. $C$ represents the number of channels, $C_{in}$ represents the number of input channels and $C_{out}$ represents the number of output channels. The first convolutional layer with $1 \times 1$ kernels is mainly used to reduce the number of input channels. Then depthwise separable convolution with $3 \times 3$ kernels is conducted. With the help of pointwise convolutional layer, features are gathered to generate the final feature maps. The feature maps are further processed with batch normalization layer and ReLU layer. CT, channel transformation [Color figure can be viewed at wileyonlinelibrary.com]

result, the number of parameters becomes $c_{in}/4 \times 3 \times 3$, which has been significantly reduced. In summary, the total number of parameters $N_{all}$ is shown as:

$$\begin{aligned} N_{all} &= c_{in} \times 1 \times 1 \times c_{in}/4 + c_{in}/4 \times 3 \times 3 \\ &+ c_{in}/4 \times 1 \times 1 \times c_{out} \\ &= c_{in}/4 \times (c_{in} + 9 + c_{out}) \end{aligned} \tag{5}$$

The number of parameters of a normal convolutional layer $N_{tra}$ is shown as follows:

$$N_{tra} = c_{in} \times 3 \times 3 \times c_{out} \tag{6}$$

Compared with (5) and (6), the ratio of the number of parameters $R$ is expressed as:

$$\begin{aligned} R &= \frac{c_{in} \times 9 \times c_{out}}{c_{in}/4 \times (c_{in} + 9 + c_{out})} \\ &= \frac{36 \times c_{out}}{c_{in} + 9 + c_{out}} \end{aligned} \tag{7}$$

When the numbers of input channel and output channel are large, $c_{in}$ and $c_{out}$ are approximate, the ratio $R$ can be approximately expressed as follows:

$$\begin{aligned} R &= \frac{36 \times c_{out}}{c_{in} + 9 + c_{out}} \\ &\approx \frac{36 \times c_{out}}{c_{in} + c_{out}} \\ &\approx \frac{36 \times c_{out}}{c_{out} + c_{out}} \\ &= 18 \end{aligned} \tag{8}$$

The above equation shows that the number of parameters of the proposed CT block is approximately 1/18 of those of normal convention layers. As a result, the designed CT module can significantly reduce the number of parameters of the network.

The efficiency of the designed CT block is discussed in the ablation studies in Section 4.2.1. Experimental results show that the proposed CT block outperforms normal convolution and group convolution layers in the proposed network.

## 3.2 | Spatial-temporal module

As mentioned above, the proposed network aims to take advantage of the 3D CNNs in spatial and time dimension. According to the experience of previous works, 3D CNNs achieve a brilliant performance when they extracts features in spatial and time dimensions,[23,35,39,40] which view the input video as a hierarchy of temporal patches and perform convolution on feature maps of different frames in the same time. Therefore, the spatial-temporal module is designed with 3D CNNs to generate a joint representation of the spatial features from four consecutive frames. As shown in Figure 1, the 3D convolutional layers perform feature extraction based on SRM features in both spatial and time dimensions. The kernels' sizes of 3D CNNs can be simplified as $t \times 3 \times 3$, where $t$ refers to time dimension. $t$ of the 3D convolution layers equal to 1, 2, and 3 separately.

## 3.3 | SRM feature extraction

As mentioned above, 3D convolution layers in spatial-temporal module are adopted to extract features in spatial and temporal dimension. Most existing 3D CNNs takes RGB frames as input.[38,42] According to the experience of steganalysis, SRM features can expose edges of the frames which may lead to better detection performance.

SRM is a set of classical steganalysis features which consist of a variety of high-pass filters including linear and nonlinear spatial high pass filters. As a result, a variety of residual frames are obtained by the submodel of SRM which are the high frequency components of the frame. SRM has been adopted in DeepFake detection and has made good performance because the residual image take more significant details from the original frames and suppress the content of the frame.

In this paper, SRM features are extracted from the four input frames and input to the spatial-temporal module instead of RGB channels. With more significant texture features, it is believed that the performance of the spatial-temporal can be improved, which helps to better utilize 3D CNNs. For our network, three $3 \times 3$ kernels are picked out from the SRM kernel sets: $H_1(x)$, $H_2(x)$, and $H_3(x)$, which are listed as follows:

$$H_1(x) = \begin{bmatrix} -\frac{1}{4} & \frac{2}{4} & -\frac{1}{4} \\ \frac{2}{4} & -\frac{4}{4} & \frac{2}{4} \\ -\frac{1}{4} & \frac{2}{4} & -\frac{1}{4} \end{bmatrix} \tag{9}$$

$$H_2(x) = \begin{bmatrix} -\frac{1}{4} & 0 & -\frac{1}{4} \\ 0 & -\frac{4}{4} & 0 \\ -\frac{1}{4} & 0 & -\frac{1}{4} \end{bmatrix} \tag{10}$$

$$H_3(x) = \begin{bmatrix} \frac{2}{4} & -\frac{1}{4} & \frac{2}{4} \\ \frac{1}{4} & -\frac{4}{4} & \frac{1}{4} \\ \frac{2}{4} & -\frac{1}{4} & \frac{2}{4} \end{bmatrix} \tag{11}$$

Ablation study is also conducted to further verify the effectiveness of the SRM features, which is introduced in Section 4.2.1. Experimental results show that SRM features can improve the performance of the spatial-temporal module.

### 3.4 | Framework

The overall framework of the proposed lightweight 3D CNN is organized as follows: as shown in Figure 1, first, four consecutive frames are taken as input and then SRM features are extracted. Then the features are input to the proposed lightweight 3D CNN.

The lightweight 3D CNN mainly consists of two modules: spatial-temporal module and CT module. The spatial-temporal module is designed to generate a joint representation of the spatial and temporal features from the input SRM features with the help of 3D CNNs.[23] The CT module is designed to extract deep level features from the output of the spatial-temporal module with least parameters, which is shown in Figure 2.

## 4 | EXPERIMENTS

In this section, experiments and further analysis are introduced. First, DeepFake data sets and implementation details are introduced in Section 4.1. Next, the experimental results are presented and the further analysis are given in Section 4.2.

### 4.1 | Data sets and implementation details

The data sets used for our experiments are FaceForensics (FF++),[25] DeepFake-TIMIT,[15] DeepFakes Detection Challenge Preview (DFDC-pre)[43] and celeb-DF.[44] Samples of the these data sets are shown in Figure 3.

FaceForensics++[25] is a large face tampering data set which is widely used in DeepFake detection. Four DeepFake methods are adopted to generate fake faces including FaceSwap,[11] DeepFakes,[13] Face2Face[12] and NeuralTextures.[42] Each type of DeepFake videos contains 1000 videos. In addition, videos are compressed with three compression rates including original compression rate videos (c0), slightly compressed rate videos (c23) and low quality videos (c40).

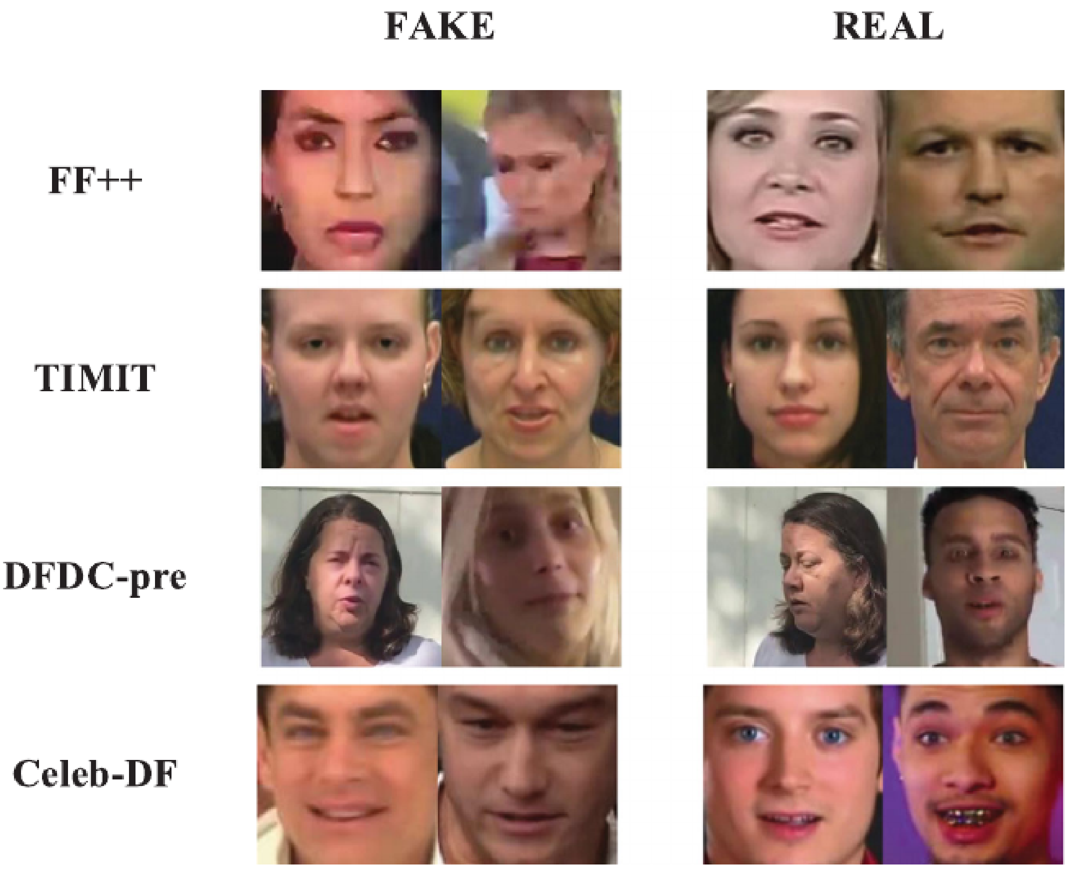

FIGURE 3 Samples of the DeepFake data sets from up to down are from FF++, DeepFake-TIMIT, DFDC-pre and Celeb-DF. Each row corresponds to a data set and two fake samples with two real samples are given, respectively [Color figure can be viewed at wileyonlinelibrary.com]

The c23 videos are close to lossless compression, whose compression quantization parameter is 23. The compression quantization parameter of c40 videos is 40.

DeepFake-TIMIT[15] is generated by the face-swapping algorithm based on VidTIMIT data set.[45] The fake videos consist of low quality (LQ) ones and high quality (HQ) ones with different resolution of the fake faces. The number of fake videos of each types is about 320.

DFDC-Preview[43] is a preview of the DFDC data set, which contains around 5000 videos. The real videos are shot by many actors of different genders, skin tones and ages with varied lighting conditions, head poses and visual diversity backgrounds. Two DeepFake methods are used to generate fake faces, which produces different qualities swaps results.

Celeb-DF[44] is a challenging DeepFake video data set, which is comprised of 590 real videos and 5639 DeepFake videos. The real videos of this data set are collected from YouTube, in which the genders, ages and ethnic groups of the celebrities are different. Using an improved DeepFake synthesis algorithm, the visual quality of the fake videos is improved.

The videos of all data sets are firstly framed. Then feature points of the faces in every frames are extracted using DLIB,[46] which helps to locate and crop the face area. The proposed network is trained by Adam[47] with a learning rate of 0.0001.

## 4.2 | Results and analysis

In this section, the experimental results and corresponding analysis are given. First, ablation studies are conducted to verify the effects of the local structure of the proposal in Section 4.2.1.

TABLE 1 The ACC(%) of networks with/without SRM feature extraction tested on Celeb-df data set

| Network | Network with SRM | Network without SRM |
|---|---|---|
| ACC (%) | **98.07** | 97.11 |

*Note*: Network with SRM feature extraction outperforms the one without SRM.

Abbreviations: ACC, accuracy; SRM, spatial rich model.

Then the experimental results and analysis on state-of-the-art data sets compared with other DeepFake detection methods are introduced in Section 4.2.2.

### 4.2.1 | Ablation studies

To verify the effect of the local structure of the proposal, ablation study is conducted where experiments are performed on Celeb-df data set.[44] In this part, the effect of SRM feature, the number of input frames and the effect of CT are evaluated.

*Effect of SRM feature*: experiments are conducted to verify the effect of SRM feature. There are two models in the experiment: one is the proposed lightweight 3D CNN with SRM features as input, the other one is the proposed lightweight 3D CNN without SRM features. The result is shown in Table 1 which is tested on Celeb-DF data set.[44] The accuracy of the network with SRM feature is approximately 1% higher than that of the network without SRM features. The result shows that the SRM features can help the proposed network get better performance.

*Effect of input frame number*: To select a more reasonable number of input frames to achieve better performance. The more frames are input, the more parameters exist. Besides, two frames are not taken into consideration because two frames may not be able to take full advantage of 3D CNNs in time dimension. Therefore, only 3, 4, 5, 6 frames are considered as input. The experimental results on the number of input frames listed in Table 2 show that the network taking four frames as input achieves best performance. Therefore, 4 seems to be a reasonable number of input sample frames which can provide enough time information and avoid too many parameters.

*Effect of the structure of CT module*: CT which is introduced above is adopted to reduce the parameters of the networks and extract higher level features. Experiments are conducted to show the performance of the CT modules with different structures. Normal convolution and group convolution which divides the input feature maps into 32 groups, are compared with CT module in the proposed network. The experimental result is shown in Table 3, where ST+NC refers to the network consisting of spatial-temporal blocks and normal convolution instead of the CT blocks, ST+GC means CT blocks are replaced by group convolution and ST+CT refers to the proposed network. The performance of the proposed CT outperforms normal

TABLE 2 The ACC(%) of networks with different numbers of input frames tested on Celeb-df data set

| Number of input frames | ACC(%) |
|---|---|
| 3 | 97.79 |
| 4 | **98.07** |
| 5 | 97.53 |
| 6 | 97.73 |

**TABLE 3** The accuracies(%) of different convolution structures tested on Celeb-df data set

| Network | ACC(%) |
|---|---|
| ST + NC | 96.23 |
| ST + GC | 97.21 |
| ST + CT | **98.07** |

*Note*: Network with depthwise separable convolution gets superior performance.

convolution and group convolution in the proposed network. Furthermore, the number of parameters of CT is much less than that of normal convolution and group convolution. Therefore, CT is the best choice of the proposed network.

### 4.2.2 | Comparison with other methods

Experiments are conducted on above state-of-the-art data sets to show the performance of the proposed networks. The proposed network are compared with two classical DeepFake detection methods including X-ception,[25] Capsule,[48] and two 3D CNN methods: I3D[42] and R3D.[38] The comparison includes two parts: the number of parameters and the detection accuracies on state-of-the-art DeepFake data sets. The number of parameters of each DeepFake detection methods are listed in Table 4 and detection accuracies tested on state-of-the-art DeepFake data sets are shown in Table 5.

As shown in Table 4, the number of parameters of the proposed network is much smaller than that of other networks. The number of parameters of the proposed network is less 1 million, approximately half of the number of the parameters of Capsule[48] which is over 1.5 million. What's more, the numbers of parameters of X-ception,[25] I3D,[42] and R3D[38] reach tens of millions. Less network parameters are not due to less network layers, but caused by the CT block. The CT block, as discussed in Section 3.1, can reduce the number of parameter of the corresponding convolution layer to 1/18 and finally contributes to the lightweight network.

Considering the conditions of the experiments, all the DeepFake detection methods take SRM features as input. The experiments are conducted on c40 samples from FF++,[25] HQ and LQ samples from DeepFake-TIMIT,[15] samples manipulated by method A from DFDC-pre[43] and all samples in Celeb-DF.[44] The experimental results are shown in Table 5. From the experimental results, the proposed method outperforms the other state-of-the-art DeepFake detection methods. The proposed lightweight 3D CNN gets accuracies over 99% on FF++[25] and DeepFake-TIMIT.[15] And the detecting accuracy on Celeb-DF[44] is 98.07%, the detecting accuracy on DFDC-pre[43] is 93.98%.

**TABLE 4** The number of parameters of the proposed network and other DeepFake detection networks

| Network | Number of parameters |
|---|---|
| X-ception[25] | 20,811,050 |
| Capsule[48] | 1,571,070 |
| I3D[42] | 12,296,594 |
| R3D[38] | 33,176,770 |
| Proposed | **851,618** |

**TABLE 5** The comparison among our proposed network and other DeepFake detection methods tested on FF++, DeepFake-TIMIT, DFDC-pre and Celeb-DF

| Methods | FF++[25] | TIMIT HQ[15] | TIMIT LQ [15] | DFDC-pre[43] | Celeb-DF[44] |
|---|---|---|---|---|---|
| X-ception[25] | 98.2 | 93.64 | 88.24 | 93.95 | 97.24 |
| Capsule[48] | 99.71 | 95.80 | 99.47 | 89.66 | 68.97 |
| I3D[42] | 97.32 | 98.71 | 97.41 | 90.98 | 96.80 |
| R3D[38] | 98.38 | 98.84 | 99.16 | 93.20 | 97.51 |
| Proposed | **99.83** | **99.28** | **99.60** | **93.98** | **98.07** |

*Note*: ACC (%) are listed.

On the one hand, considering the differences among different samples from different data sets. The satisfying performance on c40 samples in FF++[25] proves the high detection ability faced with high compression rate. The satisfying performance on DFDC-pre[43] shows the nice detection ability faced with a variety of postprocessing methods. And the satisfying performance on Celeb-DF[44] shows the effectiveness of the proposed network on high quality manipulation samples.

On the other hand, the performance of the proposed method compared with that of other methods also proves the effectiveness. Compared with the two classical DeepFake detection methods: X-ception[25] and Capsule,[48] the proposed method show better performance, which improves the effectiveness of the proposed method. Besides, compared 3D CNN-based methods: I3D[42] and R3D,[38] the proposed method gets better performance with much less parameters, which shows that the proposed method is more suitable for DeepFake detection.

The good performance of the proposed network mainly thanks to the two modules of the network: spatial-temporal module and CT module. In spatial-temporal module of the network, 3D convolution is adopted to fuse the SRM features from four input frames and process them in spatial and time dimensions, which takes full advantage of 3D convolution. And by using CT, the CT block achieves better performance compared with normal convolution and group convolution with least parameters.

## 5 | CONCLUSIONS

In this paper, a lightweight 3D CNN is proposed. CT module is designed to extract features with much fewer parameters in higher level. Serving as spatial-temporal module, 3D CNNs are adopted to fuse the spatial features in time dimension. To suppress frame content and highlight frame texture, SRM features are extracted from the input frames, which helps the spatial-temporal module achieve better performance. Experimental results show that the number of parameters of the proposed network is much less than those of other networks and the proposed network outperforms other state-of-the-art DeepFake detection methods on mainstream DeepFake data sets. As the proposed model solves the heavy deployment consumption problem while maintaining the good detection performance, it is obvious that the real application of DeepFake detection on edge devices is in near future. In future work, we will be committed to designing a more efficient and lightweight network to further play the advantages of the proposed network. More efficient 3D CNN structures will be proposed to better fuse the spatial features and more lightweight CT module will be designed to further extract features with fewer parameters.

## ACKNOWLEDGMENTS

This study is supported by the National Natural Science Foundation of China (No. U2001202, 62072480, and U1736118), the National Key R&D Program of China (No. 2019QY2202 and 2019QY(Y)0207), the Key Areas R&D Program of Guangdong (No. 2019B010136002), the Key Science Research Program of Guangzhou (No. 201804020068).

## ORCID

*Jiarui Liu* https://orcid.org/0000-0002-0002-3945
*Kaiman Zhu* https://orcid.org/0000-0001-7960-1869
*Wei Lu* https://orcid.org/0000-0002-4068-1766
*Xiangyang Luo* https://orcid.org/0000-0003-3225-4649
*Xianfeng Zhao* https://orcid.org/0000-0002-5617-8399

## REFERENCES

1. Liu X, Liu H, Lin Y. Video frame interpolation via optical flow estimation with image inpainting. *Int J Intell Syst*. 2020;35(12):2087-2102.
2. Padilla M, María J, Sanchez M, Herranz L. Video summaries generation and access via personalized delivery of multimedia presentations adapted to service and terminal. *Int J Intell Syst*. 2006;21(7):785-800.
3. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems, NIPS'14*. Cambridge, MA; 2014:2672-2680.
4. Zheng W, Yan L, Gou C, Wang F-Y. Fighting fire with fire: a spatial-frequency ensemble relation network with generative adversarial learning for adversarial image classification. *Int J Intell Syst*. 2021;36:2081-2121.
5. Chen Z, Zhu T, Xiong P, Wang C, Ren W. Privacy preservation for image data: a gan-based method. *Int J Intell Syst*. 2021;36:1-18.
6. Yan Z, Li G, Liu J. Private rank aggregation under local differential privacy. *Int J Intell Syst*. 2020;35(10):1492-1519.
7. Sun Z, Wang Y, Cai Z, Liu T, Tong X, Jiang N. A two-stage privacy protection mechanism based on blockchain in mobile crowdsourcing. *Int J Intell Syst*. 2021;36:2058-2080.
8. Chen Z, Zhu T, Xiong P, Wang C, Ren W. Privacy preservation for image data: a gan-based method. *Int J Intell Syst*. 2021;36(4):1668-1685.
9. Zhao Q, Zhao C, Cui S, Jing S, Chen Z. Privatedl: privacy-preserving collaborative deep learning against leakage from gradient sharing. *Int J Intell Syst*. 2020;35(8):1262-1279.
10. Li F, Liu Z, Li T, Ju H, Wang H, Zhou H. Privacy-aware pki model with strong forward security. Int J Intell Syst; 2020. http://doi.org/10.1002/int.22283
11. Faceswap. http://www.github.com/MarekKowalski/. Accessed September 30, 2019.
12. Thies J, Zollhöfer M, Stamminger M, Theobalt C, Nießner M. Face2face: real-time face capture and reenactment of RGB videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV; 2014:2672-2680.
13. Deepfake. http://www.github.com/deepfakes/. Accessed September 18, 2019.
14. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J. Deepfakes and beyond: a survey of face manipulation and fake detection. *Inform Fusion*. 2020;64:131-148. https://doi.org/10.1016/j.inffus.2020.06.014
15. Korshunov P, Marcel S. Deepfakes: a new threat to face recognition? assessment and detection. arXiv preprint arXiv:1812.08685; 2018.
16. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Niessner M. Faceforensics++: learning to detect manipulated facial images. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South); 2019:1-11
17. Chollet F. Xception: deep learning with depthwise separable convolutions. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI; 2017:1800-1807

18. Cozzolino D, Poggi G, Verdoliva, L. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In: *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*; 2017:159-164.
19. Afchar D, Nozick V, Yamagishi J, Echizen I. Mesonet: a compact facial video forgery detection network. In: *IEEE International Workshop on Information Forensics and Security*; 2018:1-7.
20. Li Y, Lyu S. Exposing deepfake videos by detecting face warping artifacts. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops*, Long Beach, CA; 2019:46-52.
21. Sabir E, Cheng J, Jaiswal A, AbdAlmageed W, Masi I, Natarajan P. Recurrent convolutional strategies for face manipulation detection in videos. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops*; 2019:80-87.
22. Güera D, Delp EJ. Deepfake video detection using recurrent neural networks. In: *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*; 2018:1-6.
23. Ganiyusufoglu I, Ngô LM, Savov N, Karaoglu S, Gevers T. Spatio-temporal features for generalized detection of deepfake videos. arXiv preprint arXiv:2010.11844; 2020.
24. Li L, Bao J, Zhang T, Yang H, Chen D, Wen F, Guo B. Face x-ray for more general face forgery detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA; 2020: 5000-5009.
25. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. Faceforensics++: learning to detect manipulated facial images. In: *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, Korea (South); 2019:1-11.
26. Rana MS, Sung AH. Deepfakestack: a deep ensemble-based learning technique for deepfake detectionIn: *IEEE International Conference on Cyber Security and Cloud Computing*, New York, NY; 2020:70-75.
27. LI X, YU K. A deepfakes detection technique based on two-stream network. *J Cyber Security*. 2020; 5(2): 84-91.
28. Matern F, Riess C, Stamminger M. Exploiting visual artifacts to expose deepfakes and face manipulations. In: *IEEE Winter Applications of Computer Vision Workshops (WACVW)*, Waikoloa, HI; 2019:83-92.
29. Dang H, Liu F, Stehouwer J, Liu X, Jain AK. On the detection of digital face manipulation. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA; 2020:5780-5789.
30. Cozzolino D, Thies J, Rössler A, Riess C, Nießner M, Verdoliva L. Forensictransfer: weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510; 2018.
31. Nguyen HH, Fang F, Yamagishi J, Echizen I. Multi-task learning for detecting and segmenting manipulated facial images and videos. In: *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Tampa, FL; 2019:1-8
32. Khalid H, Woo SS. Oc-fakedect: classifying deepfakes using one-class variational autoencoder. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; 2020:2794-2803.
33. Chen P, Liu J, Liang T, Zhou G, Gao H, Dai J, Han J. Fsspotter: spotting face-swapped video by spatial and temporal clues. In: *IEEE International Conference on Multimedia and Expo (ICME)*, London; 2020:1-6.
34. Xiao Y, Yin H, Zhang Y, Qi H, Zhang Y, Liu Z. A dual-stage attention-based conv-lstm network for spatio-temporal correlation and multivariate time series prediction. *Int J Intell Syst*. 2021;36:2036-2057. http://doi.org/10.1002/int.22370
35. Ji S, Xu W, Yang M, Yu K. 3D convolutional neural networks for human action recognition. *IEEE Trans Pattern Anal Mach Intell*. 2013;35(1):221-231.
36. Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M. A closer look at spatiotemporal convolutions for action recognition. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT; 2018:6450-6459.
37. Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI; 2017:4724-4733.
38. Hara K, Kataoka H, Satoh Y. Learning spatio-temporal features with 3D residual networks for action recognition. In: *IEEE International Conference on Computer Vision Workshops, ICCV Workshops*, Venice, Italy; 2017:3154-3160.
39. de Lima O, Franklin S, Basu S, Karwoski B, George A. Deepfake detection using spatiotemporal convolutional networks. arXiv preprint arXiv:2006.14749; 2020.

40. Wang Y, Dantcheva A. A video is worth more than 1000 lies. comparing 3DCNN approaches for detecting deepfakes. In: *IEEE International Conference on Automatic Face and Gesture Recognition*, Buenos Aires, Argentina; 2020:515-519.
41. Choi DH, Lee HJ, Lee S, Kim JU, Ro YM. Fake video detection with certainty-based attention network. In: *IEEE International Conference on Image Processing (ICIP)*; 2020:823-827.
42. Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI; 2017:4724-4733.
43. Dolhansky B, Howes R, Pflaum B, Baram N, Ferrer CC. The deepfake detection challenge (DFDC) preview dataset. arXiv preprint arXiv:1910.08854; 2019.
44. Li Y, Yang X, Sun P, Qi H, Lyu S. Celeb-df: a large-scale challenging dataset for deepfake forensics. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA; 2020:3207-3216
45. Sanderson c, Lovell BC. Multi-region probabilistic histograms for robust and scalable identity inference. In: *Advances in Biometrics, Third International Conference*, ICB Lecture Notes in Computer Science, Vol 5558; 2009:199-208.
46. King DE. Dlib-ml: a machine learning toolkit. *J Mach Learn Res*. 2009;10:1755-1758.
47. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: Yoshua Bengio and Yann LeCun, eds, *3rd International Conference on Learning Representations*, San Diego, CA, Conference Track Proceedings; 2015.
48. Nguyen HH, Yamagishi J, Echizen I. Echizen I. Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467; 2019.

**How to cite this article:** Liu J, Zhu K, Lu W, Luo X, Zhao X. A lightweight 3D convolutional neural network for deepfake detection. *Int J Intell Syst*. 2021;36:4990-5004. https://doi.org/10.1002/int.22499