AN EFFICIENT DEEP VIDEO MODEL FOR DEEPFAKE DETECTION

Ruipeng Sun[1,2], Ziyuan Zhao[1], Li Shen[1], Zeng Zeng[3], Yuxin Li[4], Bharadwaj Veeravalli[2], Yang Xulei[1†]

[1] Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
[2] National University of Singapore (NUS), Singapore
[3] Shanghai University (SHU), China
[4] Nanyang Technological University (NTU), Singapore

ABSTRACT

The use of deep learning technology to manipulate images and videos of people in ways that are difficult to distinguish from the real ones, known as deepfake, has become a matter of national security concern in recent years. As a result, many studies have been carried out to detect deepfake and manipulated media. Among these studies, deep video models based on convolutional neural networks have been the preferred method for detecting deepfake in videos. This study presents a novel deep video model called Sequential-Parallel Networks (SPNet) that provides efficient deepfake detection. The SPNet model consists of a simple yet innovative sequential-parallel block that first extracts spatial and temporal features sequentially, then concatenates them together in parallel. As a result, the presented SPNet possesses comparable spatio-temporal modeling abilities as most state-of-the-art deep video methods but with lower computation complexity and fewer parameters. The efficiency of the presented SPNet is demonstrated on a large-scale deepfake benchmark in terms of high recognition accuracy and low computational cost.

Index Terms— Deepfake, Deep Video Model, Sequential-Parallel Networks, Spatio-temporal Modelling.

1. INTRODUCTION

With the advent of Generative Adversarial Network (GAN) [1] and the increasing trend of its wide applications [2, 3], deepfake technology, as a key application of GAN, has received extensive attention in industry and academia. While deepfake technology has been successful in face-swapping and face reenactment, concerns about its adverse effects have increased, leading to the development of deepfake detection.

Video deepfake detection is one of the key subtasks in deepfake detection. Generally, there are two kinds of video deepfake detection methods: methods based on visual artifacts within frames and temporal features across frames. The first type of method focuses on detecting deepfake features with spatial information within each frame, while the second type of method pays attention to the temporal information across frames as well as spatial information within frames. 3D convolution [4, 5, 6] is a typical manner to extract both spatial and temporal information in the video and achieves promising performance for deepfake detection tasks [7]. However, these methods suffer from computationally intensive and time-consuming issues.

In this paper, we propose a novel network structure - sequential-parallel networks (SPNet) for efficient deepfake detection. The proposed SPNet is primarily constructed with a simple but novel sequential-parallel block that first sequentially extracts spatial and temporal features and then concatenates the features together in parallel. In such a way, to extract and preserve more spatial and temporal information in each sequential parallel block so as to improve model efficiency. The model simulates how we humans recognize deepfake features: first identify the primary object (face) in the frames and then recognize the deepfake features using combined information of spatial semantics and motion across frames. The experiments on the deepfake benchmark dataset - Deepfake Detection Challenge (DFDC) [8] demonstrate that our model achieves better performance than baselines on deepfake detection tasks.

The main contributions of this paper are three folds: Firstly, we propose a novel sequential-parallel block, then construct the lightweight SPNet model for efficient video deepfake detection. Secondly, we test SPNet on the DFDC video deepfake detection dataset and achieve better or comparable performances than baselines with fewer parameters and lower computational costs. Lastly, we release the source codes of our method to encourage further research work, which are publicly available at https://github.com/SRP009896/SPNet.git. The implementation can be easily adapted to other methods in a Plug-and-Play manner.

2. RELATED WORKS

Video deepfake detection is an essential sub-task of deepfake detection. There are two mainstream methods: spatial artifacts detection within frames and spatial-temporal artifacts detection across frames.

The spatial artifacts detection method follows the idea of image deepfake detection methods. It considers a video as a

†Corresponding author

set of images and aims to identify spatial deepfake features within frame images. Agarwal *et al.* [9] propose an algorithm to detect the inconsistency in face fusion edges. Li *et al.* [**?**] focus on the incongruity of fake face content and the rest of the part in each frame, like illumination and fidelity change. Yu *et al.* [10], and Chen *et al.* [11] detect the deepfake video frame by forensics artifacts like GAN fingerprint and PRNU. Hsu *et al.* [12] train a siamese network with real and fake video frame images to do the deepfake detection work. Wang *et al.* [**?**] consider using anomaly detection techniques to deal with video deepfake detection tasks.

The spatial-temporal artifacts detection method considers temporal deepfake features across frames as well as spatial deepfake features. Agarwal *et al.* [13] models a database for recording target people and detects anomaly behavior based on a large amount of data on the target person. Conotter [14] uses the physiological signals to detect deepfake operation. Cheng *et al.* [15] use the hearing and visual signals synergistically to check the inconsistency between audio and video. Guera *et al.* [16] focus on detecting flickering and jittering in facial regions to identify inconsistencies between frames. Another approach involves using deep learning to train a model to automatically extract deepfake features. Sabir *et al.* [17] propose the RCN model, which employs CNN to extract features within frames, followed by RNN to integrate features across frames to detect deepfakes. Nguyen *et al.* [18] introduce a time-aware pipeline that looks for inconsistencies within and between frames at a deep level in the video.

More recently, Lima *et al.* [7] adopt several 3D or pseudo-3D models for deepfake detection tasks. C3D [4] and R3D [5] use convolutional kernels in three dimensions to extract artifact features within frames and across frames at the same time. While I3D [6] uses an inflated 3D convolutional network for spatial-temporal information modeling based on convolutional network inflation. But the huge amount of computation leads to poor performance in time and space. R(2+1)D neural network [19], which uses 2D convolutional kernels with 1D convolutional kernel followed by to approximate R3D, can alleviate this situation. MCx [19] replaces parts of the 3D layer with 2D layer to decrease the computing complexity. Another improvement based on the structure of the R3D network is GST model [20], and this model aims to divide input features into two parts and use parallel convolution kernels to handle two parts at the same time. Our proposed SPNet integrates the idea of both the GST model and R2+1D model to perform efficient video deepfake detection.

3. PROPOSED DEEP VIDEO MODEL

This section describes the structure of the proposed sequential-parallel block and the architecture of the Sequential-Parallel Network (SPNet).

3.1. Sequential-parallel Block

The proposed sequential-parallel block could be viewed in two stages, a sequential stage, and a parallel stage, as illustrated in Fig. 1(c). The size of the convolution kernels is denoted by $T \times H \times W$, where T is the temporal size and H, W are the spatial height and the width of the convolution kernels, respectively.

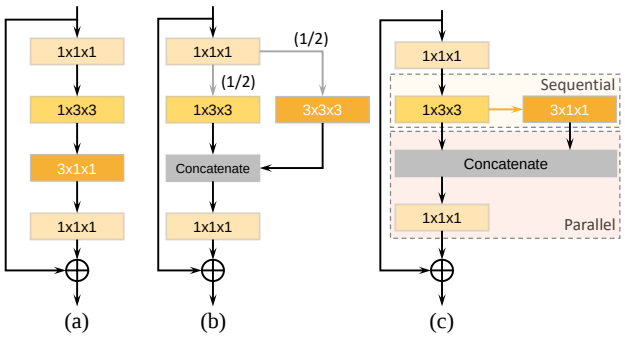

Fig. 1. Architecture of the sequential-parallel block with comparison to previous similar blocks. (a) The structure of R(2+1)D [19] block that connects spatial and temporal convolution sequentially. (b) The structure of GST [20] block that puts 2D and 3D convolution in parallel. (c) Our proposed sequential-parallel block with sequential stage and parallel stage.

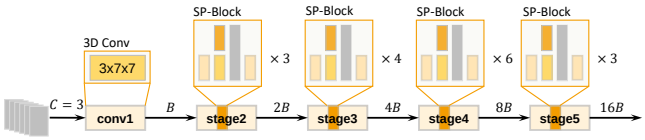

Fig. 2. **Architecture of SPNet.**

Table 1. **Instantiation of sequential-parallel block.**

Stage	Input channels	Kernel size	Output channels
-	N	$1 \times 1 \times 1$	$\gamma \times N$
Sequential	$\gamma \times N$	$1 \times 3 \times 3$	$\gamma \times N$
	$\gamma \times N$	$3 \times 1 \times 1$	$\gamma \times N$
Parallel	$2 \times \gamma \times N$	$1 \times 1 \times 1$	N

3.2. Network Architecture

Sequential stage. The sequential stage is constituted by two decomposed 2D, and 1D convolutions, where 2D convolution is performed on the spatial axes and 1D convolution is performed on the temporal axis. The spatial convolutions are to learn the spatial semantics of the input video clip. After the spatial convolutions have extracted meaningful spatial features from different frames, temporal convolution is applied to look specifically for the pattern of motions with the understanding of spatial semantics.

Parallel stage. In the R(2+1)D block, the spatio-temporal feature extracted by sequentially connected convolutions flows directly to the $1 \times 1 \times 1$ point-wise convolution. In contrast, in the sequential-parallel block, spatial features and temporal features flow in parallel into the parallel stage and are

Table 2. **Instantiation of SPNet-Bx.**

Stage	Output size	SPNet-Bx
data layer	$T \times 224^2$	-
conv1	$T \times 112^2$	$3, 3 \times 7 \times 7, x$
stage2	$T \times 56^2$	$\begin{bmatrix} d_2, 1 \times 1^2, x \\ x, 1 \times 3^2, x \\ x, 3 \times 1^2, x \\ 2x, 1 \times 1^2, 2x \end{bmatrix} \times 3$
stage3	$T \times 28^2$	$\begin{bmatrix} d_3, 1 \times 1^2, 2x \\ 2x, 1 \times 3^2, 2x \\ 2x, 3 \times 1^2, 2x \\ 4x, 1 \times 1^2, 4x \end{bmatrix} \times 4$
stage4	$T \times 14^2$	$\begin{bmatrix} d_4, 1 \times 1^2, 4x \\ 4x, 1 \times 3^2, 4x \\ 4x, 3 \times 1^2, 4x \\ 8x, 1 \times 1^2, 8x \end{bmatrix} \times 6$
stage5	$T \times 7^2$	$\begin{bmatrix} d_5, 1 \times 1^2, 8x \\ 8x, 1 \times 3^2, 8x \\ 8x, 3 \times 1^2, 8x \\ 16x, 1 \times 1^2, 16x \end{bmatrix} \times 3$
pool	$1 \times 1 \times 1$	global average pool
	# classes	fully connected layer

concatenated before a point-wise convolution is performed. Compared to merely sequential connection, the advantage of this design is the greatly enhanced interactions between spatial and temporal features by the combination of concatenation and point-wise convolution. Little extra computation is introduced in this structure.

In fact, the parallel design partly resembles the GST module proposed in [20]. As can be seen in Fig. 1(b), the input features are (optionally) divided into two parts before spatial convolution and spatio-temporal convolution are performed. However, due to the repeated extraction of spatial features in both pathways in the GST module, the sequential-parallel block is more efficient, where spatial features are reused rather than redundantly extracted.

Instantiation. The instantiation of the sequential-parallel block is presented in Table 1. As illustrated, the parameter for block-level complexity-accuracy control is introduced as γ, which is used to control the ratio between the number of input channels to the spatial convolution and the final number of output channels of the block. The value of γ is typical $1/2$ in the experiments unless otherwise specified.

Now that we have an efficient sequential-parallel block, the Sequential-Parallel Networks (SPNet) can be constructed by piling the blocks together. The architecture can be viewed in Fig. 2. As can be seen in the figure, we adopt an architecture similar to ResNet-50, which is composed of several stages, and each stage is composed of several sequential-parallel blocks. T RGB frames are sampled from the video and fed into SPNet. The first convolution layer is a simple 3D convolution, and the number of its output channel is denoted as B. In the subsequent stages, every block is the proposed sequential-parallel block. After the last sequential-parallel block, we apply a global average pooling layer and then connect the features to a fully connected layer to make the classification prediction. The number of output channels doubles after every stage that contains several sequential-parallel blocks. We name the network SPNet-Bx for $B = x$. For example, when $B = 32$, we name it SPNet-B32.

Two parameters are introduced to control the balance between model complexity and accuracy. Here the network-level complexity-accuracy control parameter is introduced and denoted as B as mentioned. Because of the described model architecture, changing B means altering the number of filters in the first block and the number of filters in all the following layers, which are based on B, with a growth rate of 2. Combined with the previous hyper-parameter γ, we could thus have a series of networks that have different levels of computation and spatio-temporal modeling ability, which can cater to different needs and application scenarios.

In our experiments, we mainly use the value of B as 24 and 32 for the network to be compact and efficient. The instantiation of SPNet can be seen in Table 2. As was mentioned before, the value of γ is $1/2$. The value of d_2, d_3, d_4, d_5 in the table represents the number of output channels of the previous sequential-parallel block. No temporal strides are used, and spatial downsampling all happens in the first block of each stage by using stride 2 in the spatial convolution layer. BatchNorm [**?**] and ReLU [21] are used after each convolution operation.

4. EXPERIMENTAL RESULTS

In this section, we evaluate and compare the performance of SPNet with some other baseline methods by measuring the test accuracy, ROC-AUC score and computation complexity.

4.1. Dataset

DFDC [8] is an open-source dataset for the Kaggle deepfake detection competition. In this paper, the DFDC-2020 dataset is used for the experiment of several algorithms. The DFDC-2020 dataset contains 119197 videos of a duration of 10 seconds or so each, with frames rates from 15 to 30 fps and resolutions from 320×320 to 3840×2160. The videos of real human faces in the dataset are filmed by 430 actors and the rest of the videos are synthesized from these videos of real faces by one of eight modification algorithms.

4.2. Experimental Setup

4.2.1. Data Pre-processing

The original videos in DFDC dataset have different frame sizes and duration and contain noise on spatial semantic information, in which case, pre-processing work is essential to unify the input of the model. More specifically, a sequence of 16 frames is captured randomly from each video in the training set and each frame is resized to 256×256. For unusual

cases, the video will be considered invalid and discarded if it contains less than 16 frames. Then, we use a 112×112 bounding box to identify and crop the recognized face region in each frame. Spatial area normalization within the channel is then applied to each frame and after that, all the frames are loaded to the tensor as the input dataset. Additionally, data augmentation is not necessary for this experiment because the synthesized videos in DFDC dataset are generated by several algorithms, which guarantees data diversity and large data volume.

Table 3. Experiment Results on DFDC Dataset

Model	Model Structure	Acc	AUC	Param Size(M)	FLOPs (G)
MC3 [19]	MC3-ResNet18	89.87	86.62	11.49	33.34
I3D [6]	I3D-ResNet18	88.95	86.34	12.29	37.22
R3D [5]	R3D-ResNet18	90.06	87.32	33.18	54.71
R2+1D [19]	R2+1D-ResNet18	90.57	87.5	31.3	40.42
GST [20]	GST-ResNet18	90.59	87.64	17.48	28.56
SPNet	SPNet-B32	**91.13**	**87.97**	14.62	24.6
SPNet	SPNet-B24	90.78	87.82	**8.09**	**17.78**

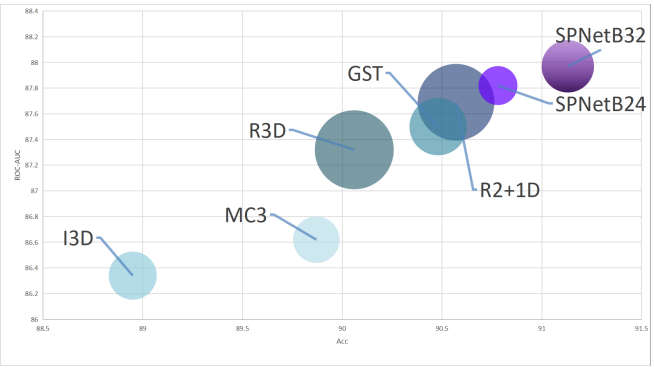

Fig. 3. Bubble Graph based on the metrics of accuracy, ROC-AUC and parameter size. x and y axes represent test accuracy and ROC-AUC, respectively and the size of bubbles represents the parameter size of models.

4.2.2. Training Method

In this experiment, the proposed SPNet and comparison algorithms are all trained from scratch and the parameters of the SPNet model are initialized by the Kaiming Initialization method. The experiment is trained on a single GPU, with batch size limited to 8. Moreover, we use SGD methods with $momentum = 0.9$ and $weight_decay = 5e - 4$ as training optimizers and weighted cross-entropy loss as criterion metrics.

4.3. Results and Analysis

We evaluate the proposed SPNet model alongside several baseline models, including MC3 [19], I3D [6], R3D [5], R2+1D [19] and GST [20]. Our analysis focuses specifically on the SPNet models with $B = 24$ and B=32, as well as the baseline R2+1D and GST models. The results, as presented in Table 3, indicate that both SPNet-B24 and SPNet-B32 achieve higher accuracy and AUC with fewer parameters and GFLOPs than other baseline models. We observe that SPNet-B24 achieves comparable accuracy and AUC to the GST model, yet with significantly fewer parameters and FLOPs. Meanwhile, SPNet-B32 achieves the highest accuracy and AUC metrics among all the models. We also provide a visualization of the experiment results in Fig 3, where the horizontal and vertical axes represent the accuracy and AUC values, respectively, while the bubble size represents the parameter size of the algorithms. The visualization also shows that our proposed SPNet-B24 and SPNet-B32 outperform other baseline methods.

The experiment results confirm that SPNet combines the strengths of R2+1D and GST models. We use a parallel channel to preserve the spatial information in the output of each block, compensating for the spatial information loss in sequential pseudo 3D kernel processing compared to the R2+1D model. Furthermore, compared to the GST model, we replace the 3D kernel branch with 2D spatial kernel and 1D temporal kernel, which decouples the temporal and spatial information and makes the optimization easier. Additionally, the model introduces more nonlinearity due to the additional ReLU between the 2D spatial kernel and 1D temporal kernel.

Table 4. Ablation study with different block size Bx

Block size	Acc	AUC	Param Size(M)	FLOPs (G)
8	88.61	85.81	4.9	11.75
16	90.18	87.16	6.81	15.03
24	90.78	87.82	8.09	17.78
32	91.13	87.97	14.62	24.6
40	91.47	88.13	26.16	30.55
48	91.58	88.25	45.79	42.81

We further investigate the effect of block size Bx on deepfake detection performance in Table 4. It can be seen that increasing Bx improved accuracy and AUC, but not as significantly as parameter size and FLOPs. Eventually, accuracy and AUC reached saturation as Bx increased further. Careful selection of Bx is crucial for balancing performance and computational resources in deepfake detection models.

5. CONCLUSION AND DISCUSSION

In this work, we present a novel sequential-parallel block based on how we humans recognize deepfake. A series of lightweight Sequential-Parallel Networks (SPNet) are constructed by stacking the blocks in groups. The proposed SPNet is compared with several baseline models using the DFDC dataset. The result shows the SPNet has a competitive performance on accuracy and ROC-AUC while significantly reducing the computational complexity.

6. REFERENCES

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," *Communications of the ACM*, 2020.

[2] Qiyu Wei, Xulei Yang, Tong Sang, Huijiao Wang, Zou Xiaofeng, Cheng Zhongyao, Zhao Ziyuan, and Zeng Zeng, "Latent vector prototypes guided conditional face synthesis," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 3898–3902.

[3] Zeng Zeng, Shenghao Zhao, Qing Da, Peisheng Qian, Tam Wai Leong, Lingyun Dai, Pär Nordlund, Nayana Prabhu, Ziyuan Zhao, and Xulei Yang, "Cyclednn-a novel deep neural network model for cetsa feature prediction cross cell lines," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2022.

[4] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[5] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh, "Learning spatio-temporal features with 3d residual networks for action recognition," in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 3154–3160.

[6] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[7] Oscar De Lima, Sean Franklin, Shreshtha Basu, Blake Karwoski, and Annet George, "Deepfake detection using spatiotemporal convolutional networks," *arXiv preprint arXiv:2006.14749*, 2020.

[8] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[9] Akshay Agarwal, Richa Singh, Mayank Vatsa, and Afzel Noore, "Swapped! digital face presentation attack detection via weighted local magnitude pattern," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 659–665.

[10] Peipeng Yu, Zhihua Xia, Jianwei Fei, and Yujiang Lu, "A survey on deepfake video detection," *Iet Biometrics*, vol. 10, no. 6, pp. 607–624, 2021.

[11] Mo Chen, Jessica Fridrich, Miroslav Goljan, and Jan Lukás, "Determining image origin and integrity using sensor noise," *IEEE Transactions on information forensics and security*, vol. 3, no. 1, pp. 74–90, 2008.

[12] Chih-Chung Hsu, Yi-Xiu Zhuang, and Chia-Yen Lee, "Deep fake image detection based on pairwise learning," *Applied Sciences*, vol. 10, no. 1, pp. 370, 2020.

[13] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li, "Protecting world leaders against deep fakes.," in *CVPR workshops*, 2019.

[14] Valentina Conotter, Ecaterina Bodnari, Giulia Boato, and Hany Farid, "Physiologically-based detection of computer generated faces in video," in *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2014, pp. 248–252.

[15] Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Tao Ye, and Liqiang Nie, "Voice-face homogeneity tells deepfake," *arXiv preprint arXiv:2203.02195*, 2022.

[16] David Güera and Edward J Delp, "Deepfake video detection using recurrent neural networks," in *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*. IEEE, 2018, pp. 1–6.

[17] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces (GUI)*, vol. 3, no. 1, pp. 80–87, 2019.

[18] Thanh Thi Nguyen, Quoc Viet Hung Nguyen, Dung Tien Nguyen, Duc Thanh Nguyen, Thien Huynh-The, Saeid Nahavandi, Thanh Tam Nguyen, Quoc-Viet Pham, and Cuong M Nguyen, "Deep learning for deepfakes creation and detection: A survey," *Computer Vision and Image Understanding*, 2022.

[19] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[20] Chenxu Luo and Alan L Yuille, "Grouped spatial-temporal aggregation for efficient action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5512–5521.

[21] Vinod Nair and Geoffrey E Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.