DeepFake Detection for Synthetic Videos by Exploiting Spatio-Temporal Information

Ishita Ghosh
ighosh4

Abstract

The recent advances in Deep Learning and Computer Vision techniques allows us to do face recognition, precise image and video classification, object recognition and many other complicated tasks. One downside of the advances are it also allows us to create sophisticated and compelling fake videos which is a major security issue off late. The major problem arises when it is used to fake someone's identity e.g. a politician saying something compromising or another application could be impersonation of identity in committing serious crime. Peddling fake news and generate fake surveillance videos are the concerns when it comes to Forensics or the crime detection experts. Existing detection methods heavily rely on the shortcomings of the generative methods that create these synthetic videos as such they focus on these intrinsic fingerprints and artifacts to distinguish deepfakes, however due to the recent advances in the generative methods, it is becoming increasingly difficult for existing state of the art image analysis methods to find such features that are spatially located in video frames. In this work we propose using the spatio-temporal features from the data to build our classifiers. Spatio-temporal features have been utilized previously for the task of video description and recognition, through experiments conducted on DFDC dataset for training pre-trained models we show that spatio-temporal features are more robust for this task.

1 Introduction

Images and videos are an integral part of the internet and many algorithms have been developed to analyze and manipulate the semantic information in them. Deepfakes are a manipulation technique that can be applied to signals like videos that can be used to swap identities of the people present in a video. Not just limited to images, Generative Adversarial Networks (GANs) [5] have been used to generate synthetic videos that are often very hard to distinguish from real ones. One popular domain for applying these techniques is in face manipulation. In particular there have been four different types of face manipulation techniques in existence: face synthesis, identity swap, expression swap and attribute manipulation. This has several security issues since the identity of a person in a video can be faked convincingly. Therefore, it is important to be able to tell when an image or a video is fake or is being generated by fraudulent parties. Detecting fake images and videos is a very active research area and as the generative methods create more realistic synthetic contents there is a necessity for deepfake detectors to improve. Traditional detection methods are heavily dependent on intrinsic fingerprints introduced by camera devices, image and color interpolation methods. FaceForensic[13] and its improved version FaceForensics++[12] proposed a benchmark for facial manipulation detection. Recently, Facebook AI released The DeepFake Detection Challenge Dataset [3] to stimulate research in this area. Wang et al. presented Fakespotter[16], which they claim offers a robust baseline for detecting fake images. These techniques can also be extended to scenarios where the input signal is a video. For example, Fernandes et al.[4] proposed a new metric for detecting fake videos.

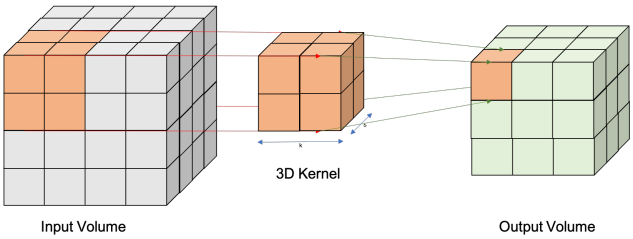

Figure 1: 3D Convolutions on Video volume

On the other hand, more and more sophisticated tools for creating fake images are coming up. StarGAN[2]

offered a way to generate new images of faces through image-to-image translation given a specific attribute. Meanwhile StyleGAN[9] showed how mapping the input to an intermediate latent space resulted in even more convincing fake images. Traditional machine learning approaches that are used to detect fake images often fail when presented with images generated by these sophisticated models[12] and the problem worsens when the dealing with HQ videos. It is observed that detecting fake video is an even harder task than fake image detection[16]. On the other hand, videos contain temporal features since each video is a sequence of images or frames and deepfakes tend to create temporal incoherence in visuals. For example, it is possible to identify artifacts introduced by affine warping performed by Deepfake generators [10]. In this work we empirically show that existing methods that rely on spatial features solely or simple sequence-based feature learning perform poorly on the latest HQ dataset. We propose that spatio-temporal features extracted using 3D CNNs Fig 1 are more robust in training the classifier for distinguishing deepfakes. In our approach we suggest using deep ResNets equipped with 3d CNNs that are pre-trained on a video dataset such as Kinetics or UCF Action Recognition when fine-tuned on the DFDC dataset shows significant improvements in classification.

2 Problem Formulation

Figure 2 represents a typical deepfake generation process where it usually uses generative method like VAE or GAN for creating the deepfake media. We model our problem as a binary classification task of detecting a video as real or fake. We propose a spatio-temporal learning machine to capture local patterns, both the spatial and temporal ones. We evaluate our approach against several baselines including models performing 2D convolutions and relying on identifying artifacts in the videos. In a single frame model based on a 2D CNN, the features extracted are mainly spatial and the result of this convolution will remain 2D over a video volume. Thus, it is more suitable to use a 3D kernel with a temporal dimension. In this work we work with a fixed temporal dimension of 16 time steps and use it as our kernel for performing 3D convolutions over the video volume.

Fig 3 shows a typical detection framework that takes a sequence of input images extracted as frames from a input video and the output label for the sequence of frames in one video as fake/real.

Formally, our task is one of binary classification. It requires us to return a trainable function f which takes a video v which is a sequence of frames $\{f_1, f_2 ... f_N\}$ and gives us a binary output (0 or 1) which tells us if the video was real or fake (0 for real and 1 for fake).

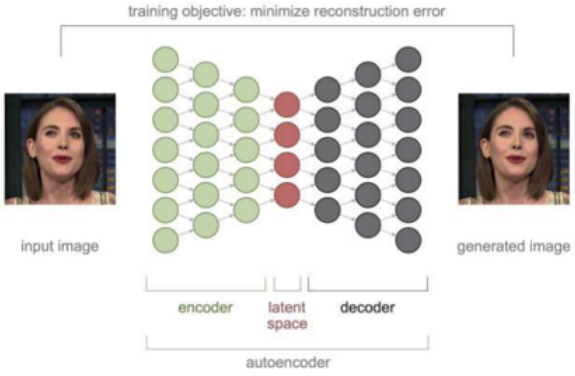

Figure 2: Deepfake generation: Typical Process

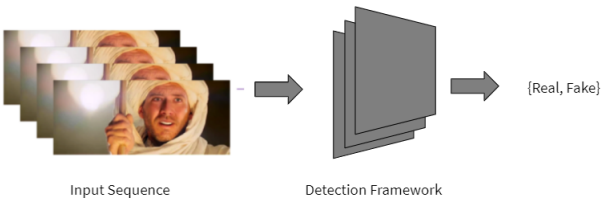

Figure 3: Detection Framework: A Typical Example

3 Deep Learning in Deep Fake Detection

Here, we discuss some of the prior work in Deepfake detection along with some popular datasets that are used as a benchmark to measure the performance of the state of the art.

3.1 Datasets:

For performing the fake detection research there are several publicly available datasets of real face images such as:

- DFDC dataset[3] - Benchmark dataset released in 2020 by Facebook to evaluate deepfake detection technology. Dataset contains train and test videos of dimension 1920 X 1080 Label for videos: FAKE or REAL Additional metadata Preview Dataset: Consists of 5000 videos based two facial altering algorithms Main Dataset: Consists of 124000 videos featuring eight facial modification algorithms.

- Celeb-DF dataset[11] - A high quality dataset containing 5,639 DeepFake videos of celebrities generated using improved synthesis process. It contains both real and DeepFake synthesized videos with similar visual quality.

In our experiments we will mainly use the DFDC

Figure 4: Celeb DF Dataset

dataset to get benchmark results for our model and compare against available baselines.

3.2 Related Work

Prior work that is relevant to our project are included below:

1. MesoNet[1]: This model uses a series of convolutional layers to perform binary classification on an input video. Their model is fast and deployable.

2. Exposing Deep-Fake Videos By Detecting Face Warping Artifacts [10]: Identifies artifacts e.g. affine warping performed by deepfake video generators

3. Deepfake Video Detection Using Recurrent Neural Networks[6]: Uses a CNN to extract features from images then passes sequence of features to RNN to identify temporal information in videos

4. SSTNet: Detecting Manipulated Faces Through Spatial, Steganalysis and Temporal Features[17]: Identifies visible tampering traces like unnatural color, shape and texture in images by extracting steganalysis features

4 Existing Methods

In order to test our hypothesis about spatio-temporal features we reviewed existing methods aimed at detecting deepfakes. 3.2.

Frame by Frame Analysis: Several existing works rely on series of 2D convolutions which are performed on the frames extracted from video. The main criterion in these works is mainly using the spatial information as features from the features or identification of noisy data from them. Exposing DeepFake Videos By Detecting Face Warping uses this approach. Artifacts[10]: The primary idea of this work is that synthetic images undergo an affine warping to match the configuration of the source's image. Mainly these warping artifacts appear around the facial region. The presence of such perturbations reveals whether it is an adversarial example or not. MesoNet [1]. is a compact, depolyable network of successive convolution layers that analyzes videos frame by frame to detect tampering. Each frame when extracted from video volume is passed through a face finding algorithm and cropped such that the network mainly uses these relevant parts of the video frame. Their best architecture achieved 95% detection accuracy on its dataset.

Spatio-Temporal Learning: CNNs have been used for performing image related tasks for a very long time, Inspired from the groundbreaking results of these CNNs in image domain, for many video related tasks such as video description (where a neural network provides the description of the activity in video), video classification (classifying a video into different activity classes such as sports, dance etc.) they were a natural choice. However, it was soon observed that these image based deep features do not perform well in training the models to learn the pattern. [7] demonstrated multiple approaches of extending the CNN in order to fuse the spatial and temporal features from videos which included kernel based modifications and time separated models. [14] demonstrated the use of a two stream architecture where the spatial extracted using 2D CNN and temporal features extracted as optical flow features were used to train the two models whose probability distribution were used to give a final combined result. [15] proposed 3D ConvNets to perform large scale supervised learning on deep architectures. Their results demonstrated that 3D convolutional deep nets worked as a good feature learning machines and a 3 X 3 X 3 kernel worked best.

4.1 Baselines

1. MesoNet[1]: In this work, the proposed model addresses the problem of detecting deepfakes primarily based on the following two methods that are commonly used to tamper facial parts: Deepfake and Face2Face. It attempts to identify video editing processes used for face tampering through a frame by frame analysis of the videos for detecting the mesoscopic features in the frames. The MesoNet dataset on which this model was trained on was extracted from the internet in this work, the resolution of the video is 384*384, where as the DFDC dataset has 1920*1080 resolution for the frames extracted from the videos. Fig 6 shows ROC/AUC of the MesoInception-4 on the MesoNet dataset and Fig 7 displays the F1 Score when the model is evaluated on MesoNet Dataset.

   We tested Meso-4 as well as MesoInception-4 architectures on the DFDC dataset. Fig 8 shows that AUC dropped to 0.5 for classifying the Real videos as well as Fake videos. Figure 9 which indicates that the model couldn't perform detection on the

Layer Name	Output size	50 - layer	101-layer
conv1	112 X 112	7 X 7, 64, stride 2	
conv2	56 X 56	3 X 3 max pool, stride 2	
		$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times3$
conv3	28 X 28	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times4$
conv4	14 X 14	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times23$
conv5	7 X 7	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times3$
	1 X 1	Average-pool, softmax	

Figure 5: ResNet Architecture

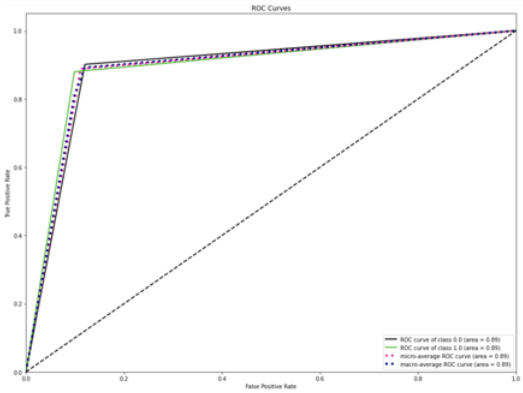

Figure 6: MesoNet ROC Curves:MesoNet Dataset

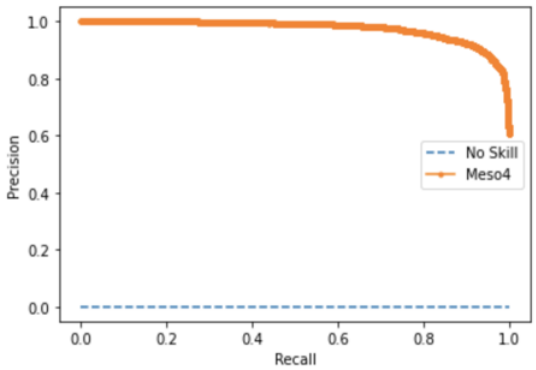

Figure 7: MesoNet: Precision vs Recall: MesoNet Dataset

HQ DFDC dataset.

2. Exposing Deep-Fake Videos By Detecting Face Warping Artifacts [10]: Identifies artifacts e.g. affine warping performed by DeepFake video generators

   This model was trained on the UADFV datset and on that dataset this model performed really well, with an AUC of 0.97. However, here we have straightway implemented the model and tested on our DFDC dataset without any tweaks or fine-tuning, mainly to evaluate the transfer learning capacity of the model.

   Fig 10 shows the result on DFDC dataset and Fig **??** displays the model performance scores in a tabular

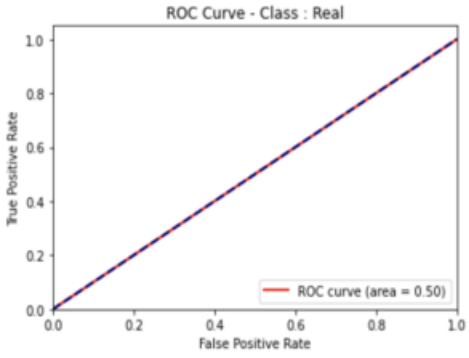

Figure 8: MesoNet: ROC for Real Class:DFDC Dataset

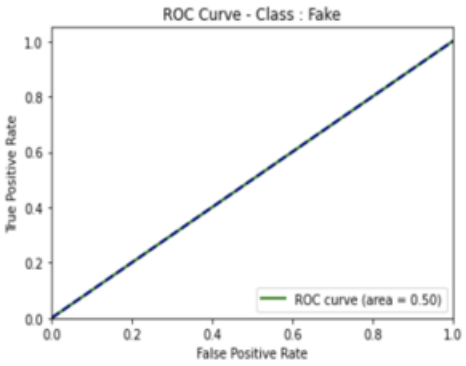

Figure 9: MesoNet: ROC for Fake Class:DFDC Dataset

format. It is clear that AUC has dropped significantly to 0.588 even though the F1 Score is 0.80.

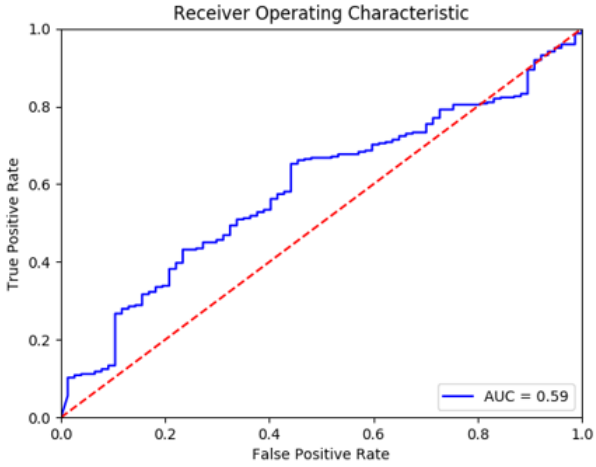

Figure 10: Detecting Face Warping Artifacts : Results on DFDC Dataset

3. 2D Resnet: As baseline we also included a 2D Resnet model pretrained on the Imagenet dataset. This model uses 2D convolutions performed over 3D video volume. As such the model's performance allows us to support our hypothesis of using 3D convolutions over video volume. The model follows the typical architecture of Resnet50 with pretrained weights from the ImageNet, we fine tuned the upper layers of this model using the DFDC dataset and obtained an 84% accuracy on the test set.

5 Implementation & Evaluation

We propose using 3D ResNets for extracting spatio-temporal information from the video data to train the network. Our main hypothesis is that features extracted using 3D convolutions identify more robust representations in the data distribution. We evaluate our approach against the baselines to determine a more robust classifier. CNNs have been used for large scale video classification [8] by extending their connectivity in time domain.

We implement two 3D ResNet models: ResNet50 and ResNet101. figure 5 Both the models are pre-trained on the Kinetics dataset. We fine tune the upper layers of these models using the DFDC dataset, this allows us faster training, reduced GPU usage and more generalizable model. The use of 3D CNN is motivated from the works of [14] which show that spatio-temporal features that were obtained by pre-trained models fine tuned over the problem actually outperformed existing methods.Figure 11

Model	3-fold Accuracy
Soomro et al [22]	43.9%
Feature Histograms + Neural Net	59.0%
Train from scratch	41.3%
Fine-tune top layer	64.1%
Fine-tune top 3 layers	**65.4%**
Fine-tune all layers	62.2%

Figure 11: Transfer Learning improves performance

We used the DFDC dataset and performed standard data augmentation tasks which include: Random cropping, Brightness adjustment, Random Flipping and Normalization of the data. For normalization we used the mean and standard deviations of the Kinetics dataset. These augmentation techniques allowed us to improve the diversity of data where as normalization allow us to center the data. The baseline and all the ResNet models (2D and 3D) were trained and tested using NVIDIA GeForce GTX 1080 Ti GPU.

Implementation: We used Pytorch implementations of the backbone models, due to memory and time constraints prevented us from testing more deeper architectures such as ResNet200, however we were able to demonstrate favorable results with ResNet101. In the both the model we have 3 X 3 X 3 kernel that processes temporal information across 16 consecutive frames simultaneously. All the pre-trained weights for the ResNets based on ImageNet and Kinetics-400 dataset were publicly available. We trained the ResNet models using Adam and using Cross Entropy Loss as the optimization criterion. For evaluation we performed an analysis on the dataset and observed an imbalance in the 2 classes which more favored towards fake class. Although our best model outperforms the baselines in all the criteria, we suggest considering the F1 score as evaluation metric for ranking the classifiers due to the imbalanced

Model	Test Accuracies (%)	Precision	Recall	Dice Coefficient / F1
Face Warping Artifacts Identification	69%	0.82	0.79	0.80
MesoNeT: Detect Facial Forgeries	78%	0.84	0.88	0.86
Single Frame ST feature Fusion	78%	0.79	0.90	0.85
3D ResNet50 pretrained on Kinetics	84%	0.82	0.90	0.86
3D ResNet101 pretrained on Kinetics	**85%**	**0.86**	**0.98**	**0.92**

Figure 12: Model Performance Comparison on DFDC Dataset

nature of the dataset. Figure 12 shows the summary of our results on the DFDC dataset for different architectures. These results demonstrate that ResNet101 exhibits a more robust performance compared to the other models. The F1 score which is a harmonic mean of precision and recall indicates that the 3D ResNet101 outperforms all the classifiers. In terms of test accuracy we can see that the 3D Resnets outperform the models based on 2D convolutions by a very large margin indicating that temporal information played a vital role in the decision of the classifiers on this problem.

6 Conclusion

In this work we attempt to address the problem of classifying deepfake videos by learning spatio-temporal features. We based our hypothesis on the observation that existing detection models either spatial or sequential learn the spatial features of the data however, when working on videos especially HQ videos, the temporal information also plays an important role and must be included for training the classifiers. We demonstrated that models based on 3D CNNs are more capable in extracting and fusing the spatial and temporal information from these datasets and deep neural networks that use 3D convolutions as feature extractors significantly outperform the models that use 2D convolutions for extracting spatial features.

7 Scope

- As a further extension of this work we intend to test with the generalizability of our models. Although the models have been trained based on transfer learning from different relevant datasets we intend to test them using other HQ datasets.
- We tested 3D ResNets of 50 layers and 101 layers and we intend to test more deeper architectures including a 152 layered and 200 layered which are more computationally costly. Training these deep neural nets from scratch have shown to perform poorly so we will be again using the pretrained weights from Kinetics-400 dataset.
- In our work we tackled this problem as a binary classification problem, this problem can also be modeled from a image segmentation perspective where frame segments could be identified as perturbed or not perturbed. A possible direction could be using robust segmentation models combined with spatio-temporal feature learning methods.

References

[1] AFCHAR, D., NOZICK, V., YAMAGISHI, J., AND ECHIZEN, I. Mesonet: a compact facial video forgery detection network. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)* (2018), 1–7.

[2] CHOI, Y., CHOI, M.-J., KIM, M., HA, J.-W., KIM, S., AND CHOO, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 8789–8797.

[3] DOLHANSKY, B., BITTON, J., PFLAUM, B., LU, J., HOWES, R., WANG, M., AND FERRER, C. C. The deepfake detection challenge dataset, 2020.

[4] FERNANDES, S., RAJ, S., EWETZ, R., PANNU, J. S., KUMAR JHA, S., ORTIZ, E., VINTILA, I., AND SALTER, M. Detecting deepfake videos using attribution-based confidence metric. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2020), pp. 1250–1259.

[5] GOODFELLOW, I. J., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A., AND BENGIO, Y. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (Cambridge, MA, USA, 2014), NIPS'14, MIT Press, p. 2672–2680.

[6] GUERA, D., AND DELP, E. Deepfake video detection using recurrent neural networks. *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (2018), 1–6.

[7] KARPATHY, A., TODERICI, G., SHETTY, S., LEUNG, T., SUKTHANKAR, R., AND FEI-FEI, L. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (2014), pp. 1725–1732.

[8] KARPATHY, A., TODERICI, G., SHETTY, S., LEUNG, T., SUKTHANKAR, R., AND FEI-FEI, L. Large-scale video classification with convolutional neural networks. In *CVPR* (2014).

[9] KARRAS, T., LAINE, S., AND AILA, T. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 4396–4405.

[10] LI, Y., AND LYU, S. Exposing deepfake videos by detecting face warping artifacts. In *CVPR Workshops* (2019).

[11] LI, Y., YANG, X., SUN, P., QI, H., AND LYU, S. Celeb-df: A large-scale challenging dataset for deepfake forensics. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 3204–3213.

[12] RÖSSLER, A., COZZOLINO, D., VERDOLIVA, L., RIESS, C., THIES, J., AND NIESSNER, M. Faceforensics++: Learning to detect manipulated facial images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 1–11.

[13] RÖSSLER, A., COZZOLINO, D., VERDOLIVA, L., RIESS, C., THIES, J., AND NIESSNER, M. Faceforensics: A large-scale video dataset for forgery detection in human faces.

[14] SIMONYAN, K., AND ZISSERMAN, A. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems 27* (2014), 568–576.

[15] TRAN, D., BOURDEV, L., FERGUS, R., TORRESANI, L., AND PALURI, M. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (2015), pp. 4489–4497.

[16] WANG, R., JUEFEI-XU, F., MA, L., XIE, X., HUANG, Y., WANG, J., AND LIU, Y. Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20* (7 2020), C. Bessiere, Ed., International Joint Conferences on Artificial Intelligence Organization, pp. 3444–3451. Main track.

[17] WU, X., XIE, Z., GAO, Y.-T., AND XIAO, Y. Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020), 2952–2956.