Towards Robust and Generalized DeepFake Detection

Siddharth Yadav
Department of Applied Mathematics,
Delhi Technological University,
New Delhi, India
thesiddharthyadav1@gmail.com

Sahithi Bommareddy
Department of Software Engineering,
Delhi Technological University,
New Delhi, India
sahithibommareddy@gmail.com

Dinesh Kumar Vishwakarma
Department of Information Technology,
Delhi Technological University,
New Delhi, India
dvishwakarma@gmail.com

2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT) | 978-1-6654-5262-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCCNT54827.2022.9984553

***Abstract*— Images that are manipulated are prevalent and are on the spike because of the advancement in deep convolutional neural networks (CNNs) techniques. There have been several concerns regarding the advent spread of false information. There exists a need for a reliable and robust method to detect such fake images. In this paper, analysis was done using the architecture SlowFast in detecting manipulated videos. This paper focuses on detecting DeepFake videos under three distinct scenarios, which are (i) all manipulation detection, (ii) single manipulation detection, and then (iii) cross manipulation detection used to test the veracity of the videos. The manipulation methods and designing algorithms to categorize such unknown manipulation techniques were used.**

Keywords*— Deepfakes, Deepfake Detection, Media forensics, Computer Vision, SOTA(*State-Of-The-Art*), Facial Manipulation Detection

I. Introduction

DeepFakes are the techniques created to impose a face on the original face of the target person. It is to make the video realistically so that the audience will not be able to find the discrepancy. FaceSwap is the category from which deepfake techniques originated. In this, the lip sync and the facial expressions are synced to make the synthesized content readily believable. There is a type of deepfake called puppet deepfake. In this, the target individual and the animations of facial features' expressions are followed by gestures and eye movements. DeepFake detection is a technique that first originated in 2017.

In the Computer Vision field, generating a deepfake is a novel research area. If generating a deep fake attracts many researchers, ethical deepfake detection should also be encouraged. What we see is not always the fundamental idea. There can be any fraud around us without getting the slightest hint. Fake news is ubiquitous. Social media has made it possible to find so much related information, but such incorrect information is also pretty standard. DeepFake detection is also capable of fostering the interventions of democratic institutions as well. Because of this, many deepfake videos are released during important events like elections, making it as if the politician is speaking evasively [1].

Contributions: As explained above, this work has motivated us to compare the SOTA video-based techniques and detect the DeepFakes. We also believe that the current state-of-the-art techniques [2], [3], [4], [5], [6], [7], [8], [9] omit a pertinent clue which is by only using the spatial information for an investigation. There have also been models that include difficulties in safeguarding the appearance of the generated videos and maintaining motion consistency[10]. The SlowFast networks use the 3DCNN architecture and outperform the image-based state-of-the-art techniques. We also plan to show that such trained models are known to detect even if we manipulate the techniques and are generalized to work on the methods defined outside the training set. We then use this evaluation technique and train the datasets, so they do not tamper with the manipulation techniques.

II. Related Work

FaceForensics++ is the most commonly used dataset for facial manipulation [11]. Wang et al. reviewed all images, graphs, and text manipulation techniques. In this, there is comprehensive research for [12].

- *Image and Video Generation*: Several approaches enabled the Generative Adversarial Networks (GANs), which focus on identity and the expressions of the face. [13], [14], [15].
- *Detection of DeepFake:* Several approaches are based on manipulation techniques and work on the videos and audio videos. Some approaches based on videos are said to perform better than image-based papers. This type is mainly used for detecting attacks. Fridrich et al. used steganalysis and performed a facial re-enactment video detection on the dataset [2]. In [6], they used an algorithm to visualize the artifacts that are generated by computers to detect the manipulated images by the computer. Therefore, the work in this field is well documented.
- *Adversarial Detection*: In this, a situation with a minute feature makes the models give false results. In this, the model tends to become weakened. So Wang et al. presented that the manipulated video was detected as real, which is a false prediction.

In [3], Rossler et al. took the FaceForensics++ dataset as well and created a model to detect the adversarial examples. Hermandez-Ortega et al. used deepfake detection, which is based on the estimation of the heart rate. He used the skin color change, which was very subtle and revealed the presence of the human skin. He used the Convolutional Attention network [16], which used facial and temporal information [17].

In the DPNet, the predictions between the similarities helped access the deep fake detections. This was implemented on unseen testing datasets, the DeepFakeDetection and Face Forensics++, a standard dataset for DeepFake detection. Through this, they obtained dynamic visual explanations and case-based reasoning. The LQ dataset presented an AUC of 90.91. This could have been further extended on the Celeb-DF dataset as well. They used the ablation study using the various diversity regulation and the series of flow frames [18].

Keeping all the future work in the industry that has been done till now in mind, our work focuses on 3D CNN mechanism and incorporates temporal and spatial information of the video. 3D CNN incorporates another dimension which is time in comparison to the 2D CNN.

III. SOTA DeepFake Detection methods

This section describes the SOTA method for DeepFake detection that has been used for the comparative study in this paper. The model uses ResNet base architecture, namely SlowFast (ResNet50 and ResNet101) was implemented. The networks were trained on the scale for human action and then on the dataset of the Kinetics-400, which is an action-based dataset[49]. It has experimentally been found that models pretrained on Kinetics-400 perform better[19]. Further, we assigned the pretrained weights in the initial phase in the modified layers of the neural network model. Then, there was a fine-tuning of the networks which were then used for the FaceForensics++ dataset.

In the SlowFast model, there is one model with high definition. It also uses a simultaneously slow structure and a fast path to analyze the content carefully and then runs it on a slow pathway. This technique is used to make the content dynamic. This is the technique in which $\boldsymbol{4/5th}$ of the cells operate at a low frequency and observe the minute details. In the other $\boldsymbol{1/5th}$ of the cells, they operate at a higher frequency and change very quickly to whatever is added to it.

SlowFast works simultaneously by capturing frames simultaneously and then using 3D operations on them. In this, the 3D ResNet Model has slow and fast pathways. In the slow pathway, there is a temporal stride, meaning the number of frames skipped and is usually set to 16, and for the fast pathways, it is usually 8. The data from the fast pathways is then input into the slow pathways, which forms a type of connection as depicted in FIGURE 1. This allows the slow pathway to know the result from the fast pathways and move on. The SlowFast model performs a series of data transformation techniques on the fast pathways and then concatenates the slow pathway into the fast pathway.

There are three techniques of the data transformation involved:

- *Time to channel*: In this, there is restructuring and transposing of the frames
- *Time strode sampling*: In this, there is a simple sampling technique that samples all of the frames
- *Time stridden convolution* is a technique that performs a 3D convolution

The research paper [20] uses a two-way path that utilizes the SlowFast Model. In this, the semantic data captures the low frame rates and gives the slow refreshing speed. This is responsible for changing the movement. Because of its lightweight computation, 20% of its utilization is total. These two pathways are combined using lateral connections.

SlowFast was previously used on Kinetics 400 and the AVA dataset in the previous works. The Kinetics 400 dataset included a 10-second sequence of YouTube videos, and the AVA dataset included a 15-minute YouTube video with actions.

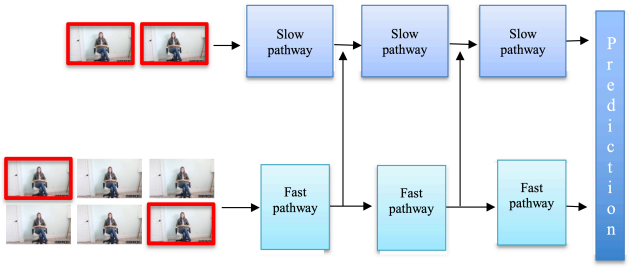

FIGURE 1. A HIGH-LEVEL ILLUSTRATION OF THE SLOWFAST NETWORK

IV. Experimental Setup

In this section, we summarise the dataset and the experimentation setup.

A. Dataset

The FaceForensics++ dataset has 1000 subjects who are talking and show slight movements [3]. It is represented in 1000 real videos, and 4 x 1000 adversarial examples. The FF++ dataset follows the four manipulation formats:

a. *faceswap*: A graphic approach replaces the frame region from source to target. This uses a model that deploys blended shapes and fits them into the transferred face.

b. *faceswap*: Using FakeApp and FaceSwap GitHub are the originating sources of DeepFakes in general

c. *Face2Face*: The facial recognition system [21] known as Face2Face follows the technique of transferring the region of the face from the source to the target. However, it does not compromise the identity of the person it derives from.

d. *NeuralTextures*: Neural Textures [22] is a primitive graphics technique that focuses on the dimension and then stores it in a learned feature vector per pixel.

Rossler et al. used FF++, a dataset, to compare the deep neural and hand-crafted networks. In these examples we used adversarial natured images and videos [3]. He had done it on raw data, low-quality data, and high-quality data, which are termed as HQ and LQ individually. In the LQ setting, they first used all the manipulation methods in the training phase and then secondly individual training on the methods separately. Then they used the first training set in a more challenging setup. In this, the raw data is used and is obtained by using the XceptionNet model. And then, the model detected high quality and low detection rates [22].

The dataset used state-of-the-art methods to trim the videos and check for face manipulation. For the four SOTA manipulation methods, 1000 videos were used, and the videos were collected. After this, the YouTube videos had to be tested out in which they were seen by observing them manually and then used to select them as the input. Out of the 1000 videos, they used 914 of them. They used the approach of FaceSwap and Face2Face as well. For the learning based approach, they used NeuralTextures and DeepFakes as well. Therefore they used four models to make the manipulations possible, which are explained below and shown in FIGURE 2.

FIGURE 2. DIFFERENT TYPES OF FACIAL MANIPULATIONS. FROM LEFT TO RIGHT: ORIGINAL SOURCE (LARGE) AND TARGET (SMALL) IMAGES, DEEPFAKES , FACE2FACE, FACESWAP , NEURALTEXTURES [12]

B. Experiment Resources and Pipeline

Pytorch was used to implement the models in this paper. 2 NVIDIA TU102 [TITAN RTX] GPUs were used to make an end-to-end training of the models. Adam optimizer was used as the learning algorithm and learning rates were set to a constant value of $\mathbf{1e^{-3}}$ across all our experiments. Then the cost function used was Cross-Entropy loss function to train the model. In the training, the models require model-pipelines with videos of 64 frames and of spatial dimensions $\mathbf{224 \times 224}$ as the input. The training pipeline then uses Temporal Subsampling, Normalisation, Scaling and Cropping according to the model architecture in consideration at the time. Then the detection of each face using the Dual Shot Face Detector is done [23]. After this, the cropped region area is increased by a factor of 1.3 . This helps to uniformly get the sides of the image so that more information can be gained by the model.

C. Experiment Implementation

There was an experiment of the above-listed architecture, SlowFast, on all four manipulation techniques. Their primary aim is to train and detect the single manipulation techniques, separate techniques for each model, and also the cross manipulation techniques. After this, the data was split into training, testing, and validation according to the guidelines of a protocol in the FaceForensics++ dataset.

1) All manipulation techniques

Primarily, the TCR (True Classification Rate) method was used, and then the five models of the CNNs were analyzed and then compared to the algorithms which have image forgery detection. Then the SOTA methods like XceptionNet and these learning methods are used to base a structure of the generic manipulation. A SOTA structure is given by Abhijit et al. and Wang et al. [12], [19]. We then use the experiment results and observe that the video-based algorithm achieved better accuracy than the image-based algorithm. The ResNet model and the nonlocal block show the best metrics in the experiments performed before. There is also an unbalanced sample problem as there are four times of fake videos than the number of original videos; we can reduce the effects of unbalanced data by using weighted cross-entropy loss. Related results are depicted in TABLE 1.

TABLE 1. DETECTION OF ALL FOUR MANIPULATION METHODS, LQ. TRUE CLASSIFICATION RATES ON NEURALTEXTURES (NT), FACE2FACE (F2F), FACESWAP (FS), DEEPFAKES (DF)

Algorithm	Train and Test	TCR
Rahmouniet al. [6]	F2F, NT, DF, FS	61.18
Bayar and Stamm [5]	F2F, NT, DF, FS	66.84
Steg. Features + SVM [2]	F2F, NT, DF, FS	55.98
MesoNet [7]	F2F, NT, DF, FS	70.47
Cozzolino et al. [4]	F2F, NT, DF, FS	58.69
XceptionNet [22]	F2F, NT, DF, FS	81.0
3D ResNet [12]	F2F, NT, DF, FS	83.86
3D ResNet (with non-local) [19]	F2F, NT, DF, FS	86.72
3D ResNet (with SE) [19]	F2F, NT, DF, FS	80.0
3D ResNeXt [12]	F2F, NT, DF, FS	85.14
I3D [12]	F2F, NT, DF, FS	87.43
Timesformer [19]	F2F, NT, DF, FS	82.3
SlowFast (modified)	F2F, NT, DF, FS	**87.5**

2) Single Manipulation Techniques

The performance measures of all the models that are trained and tested on data that is created using single manipulation techniques are then computed. The values obtained are used to report the TCRs in TABLE 2. As video-based algorithms can also reduce the data size, which helped to pertain to the videos of a single manipulation technique better than the image-based algorithms. The experiments also show that the detection approaches are most confused by the NeuralTextures approach. It can be observed that the NeuralTextures train the model, which was unique for every video, this results in a more significant artifact variation. The DeepFakes is similar as it trains to a model per video, in which a fixed processing technique in the post order of pipeline is used, and that has consistent artifacts.

TABLE 2. DETECTION OF EACH MANIPULATION METHOD INDIVIDUALLY, LQ. ACCURATE CLASSIFICATION RATES REPORTED ON NEURALTEXTURES (NT), FACE2FACE (F2F), FACESWAP (FS), DEEPFAKES (DF)

Algorithm	NT	F2F	DF	FS
Rahmouniet al. [6]	60.07	64.23	85.45	56.31
Bayar and Stamm [5]	70.67	73.72	84.55	82.52
Steg. Features + SVM [2]	63.33	73.72	73.64	68.93
MesoNet [7]	40.67	56.20	87.27	61.17

Algorithm	NT	F2F	DF	FS
Cozzolino et al. [4]	78.00	67.88	85.45	73.79
XceptionNet [22]	80.67	86.86	96.36	90.29
3D ResNet [12]	73.5	89.6	91.81	88.75
3D ResNet (with non-local) [19]	78.29	91.25	95.16	94.11
3D ResNet (with SE) [19]	66.25	77.00	81.70	75.90
3D ResNeXt [12]	80.5	86.06	93.36	92.5
I3D [12]	80.5	90.27	95.13	92.25
Timesformer [19]	67.5	73.9	90.7	82.1
SlowFast (modified)	**82.75**	**93**	**97**	**95.75**

Studies to determine the optimal depth of the SlowFast network are performed, for which the results for two SlowFast architectures with different depths, SlowFast50 (with ResNet50 as base architecture) and SlowFast101 (with Resnet101 as base architecture), are noted. The final results are summarised in TABLE 6. Comparing the results of Table II and TABLE 6, we can see that both of the SlowFast networks, the SlowFast50, and the SlowFast101, outperform all the other architectures. It can be summarised from TABLE 6 that for all the scenarios, the SlowFast50 performs almost similar or even slightly better in some scenarios than SlowFast101. It can be observed from TABLE 6 that SlowFast50 has 45% less number of parameters than SlowFast101 and hence its cost (measured in FLOPs) per spacetime view is also low by 48%. Hence, the SlowFast50 architecture is more lightweight and computationally effective than the SlowFast101 architecture. Hence it can be considered that SlowFast50 is the SlowFast model used in this research paper for the rest of the experiments.

As it can be seen, this SlowFast model outperforms all other models in the experiment and other SOTA models when reporting TCR. The SlowFast model was compared to the previous SOTA methods from Afshar et al. [7], Rossler et al. [3], and Sabir et al. [24] concerning AUC and Accuracy in TABLE 3 and TABLE 4.

TABLE 3. DETECTION OF EACH MANIPULATION METHOD INDIVIDUALLY, LQ. AUC (AREA UNDER THE ROC CURVE) REPORTED ON NEURALTEXTURES (NT), FACE2FACE (F2F), FACESWAP (FS), DEEPFAKES (DF)

Algorithm	DF	F2F	FS	NT
Sabir [24]	96.9	94.35	96.3	-
SlowFast (modified)	**98.96**	**97.65**	**98.13**	-

TABLE 4. DETECTION OF EACH MANIPULATION METHOD INDIVIDUALLY, LQ. ACCURACY REPORTED NEURALTEXTURES (NT), FACE2FACE (F2F), FACESWAP (FS), DEEPFAKES (DF)

Algorithm	DF	F2F	FS	NT
Meso-4 [7]	89.1	83.2	-	-
Mesoinception-4 [7]	91.7	81.3	-	-
Rossler [25]	94	-	93	-
SlowFast (modified)	**97**	**93**	**95.75**	-

For all approaches, in both single-manipulation-techniques and all-manipulation-techniques compared using metrics of TRC, AUC, and Accuracy, the SlowFast model performed more with quality than the other existing methods. The Precision, Recall, F1 Score, and other metrics for the SlowFast model are reported in TABLE 5.

TABLE 5. DETECTION OF EACH MANIPULATION METHOD INDIVIDUALLY ON DIFFERENT METRICS, LQ REPORTED ON NEURALTEXTURES (NT), FACE2FACE (F2F), FACESWAP (FS), DEEPFAKES (DF)

Metric	DF	F2F	FS	NT
AUC	98.96	97.65	98.13	88.91
Precision	96.5	91.22	92.38	84.70
Recall	96.5	93.5	97	77.5
F1 Score	96.5	92.35	94.63	80.94

3) Cross Manipulation Techniques

In this experiment, the model is trained on 3 of the manipulation techniques. Then with that particular architecture, the model was tested on the remaining manipulation technique. The results of this particular experiment are given in TABLE 7. This is the most challenging of the three experiments; however, it is the closest to reality. As we used the weighted cross-entropy for the first experiment, the weighted cross-entropy was used in this as well. As the testing videos have been manipulated with a different technique, it is implausible to say that the trained deepfake model has a recollection of the manipulated techniques. In the Face2Face and FaceSwap techniques, some approaches were based on graphics. Unlike the learning-based approaches pointing to DeepFakes and NeutralTextures. Though these are learning-based, the FaceSwap makes sure it transposes the facial part as the target image and then uses advanced blending. In this, there are many color correction algorithms to dominate the source over the target as well seamlessly. Therefore, there is a difference between the FaceSwap technique and the learning-based approaches. However, there is a good blend in the images. Human detection techniques of manipulation usually affect the FaceSwap and Deepfakes. They are usually more challenged using the Face2Face and ultimately NeuralTextures as well. The ranking corresponds to those CNN results discussed in experiment 2. Therefore, it can be said that the SlowFast model outperformed all the other state-of-art networks in this particular scenario.

TABLE 6. COMPARISON OF PERFORMANCES OF DIFFERENT SLOWFAST ARCHITECTURES ON EACH MANIPULATION METHOD INDIVIDUALLY, LQ. ACCURATE

Algorithm	Depth	Params	GFLOPs×views	DF	F2F	FS	NT	All
SlowFast	ResNet50	~34 M	~65x30	97	93	95.75	82.75	87.5
SlowFast	ResNet101	~62 M	~127x30	97	92.25	95.25	82.75	86.9

CLASSIFICATION RATES REPORTED ON NEURALTEXTURES (NT), FACE2FACE (F2F), FACESWAP (FS), DEEPFAKES (DF)

TABLE 7. DETECTION OF CROSS-MANIPULATION METHODS, LQ. ACCURATE CLASSIFICATION RATES REPORTED ON NEURALTEXTURES (NT), FACE2FACE (F2F), FACESWAP (FS), DEEPFAKES (DF)

Methods	Train Dataset	FS, NT, F2F	FS, NT, DF	F2F, DF, NT	FS, F2F, DF
	Test Dataset	DF	F2F	FS	NT
3D ResNet [12]		75.36	74.29	59.64	64.29
3D ResNet (with SE) [19]		52.5	53.5	53.5	55.35
3D ResNet (with non-local) [19]		77.8	73.1	65.71	65.2
3D ResNeXt [12]		75.00	70.71	57.14	68.57
I3D [12]		72.50	68.93	55.71	66.79
Timesformer [19]		65.0	62.5	53.6	60.0
SlowFast (modified)		**91.95**	**95.5**	**85.25**	**87.15**

V. CONCLUSION

In the research paper, deepfake detection was compared using some of the SOTA methods using different manipulation techniques. It tested out the SlowFast technique, which was adapted using action recognition. Then this dataset was used for pretraining the techniques, and then the method was generalized to detection of DeepFakes. The experiments showed exceptional results in which the previous-based image-based forgery detection algorithms outperformed. The reason is that networks generally lack the ability to transfer learned knowledge from the trained manipulation method to the tested manipulation method. Therefore, machine learning models might exhibit unpredictable behavior.

The limitation of this research paper is that the dataset used is a little older compared to the DFDC Preview, which is a more recent database and works on a more generalized principle. The current technique is just based on images and video and does incorporate audio. Hence, this is a limitation we plan to eliminate in the future.

In the future, a focus will be made on the additional deepfake manipulation techniques. A plan was developed as a deepfake approach based on pixel-level generated noise. It is also believed to target some of the manipulation algorithms as well.

REFERENCES

[1] J. Silbey and W. Hartzog, "Maryland Law Review The Upside of Deep Fakes," 2019. [Online]. Available: https://publicintegrity.org/federal-politics/the-citizens-united-decision-

[2] J. Fridrich and J. Kodovský, "Rich Models for Steganalysis of Digital Images," 2012.

[3] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," Jan. 2019, [Online]. Available: http://arxiv.org/abs/1901.08971

[4] D. Cozzolino, G. Poggi, and L. Verdoliva, "Recasting Residual-based Local Descriptors as Convolutional Neural Networks," in *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, Jun. 2017, pp. 159–164. doi: 10.1145/3082031.3083247.

[5] B. Bayar and M. C. Stamm, "A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer," in *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, Jun. 2016, pp. 5–10. doi: 10.1145/2909827.2930786.

[6] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Distinguishing computer graphics from natural images using convolution neural networks," in *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, Dec. 2017, pp. 1–6. doi: 10.1109/WIFS.2017.8267647.

[7] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," Sep. 2018, doi: 10.1109/WIFS.2018.8630761.

[8] A. Ajoy, C. U. Mahindrakar, D. Gowrish, and V. A, "DeepFake Detection using a frame based approach involving CNN," in *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, Sep. 2021, pp. 1329–1333. doi: 10.1109/ICIRCA51532.2021.9544734.

[9] N. Kumar, P. P, V. Nirney, and G. V, "Deepfake Image Detection using CNNs and Transfer Learning," in *2021 International Conference on Computing, Communication and Green Engineering (CCGE)*, Sep. 2021, pp. 1–6. doi: 10.1109/CCGE50943.2021.9776410.

[10] Y. Wang, P. Bilinski, F. Bremond, and A. Dantcheva, "G3AN: Disentangling Appearance and

Motion for Video Generation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 5263–5272. doi: 10.1109/CVPR42600.2020.00531.

[11] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. Jain, "On the Detection of Digital Face Manipulation," Oct. 2019, [Online]. Available: http://arxiv.org/abs/1910.01717

[12] Y. Wang and A. Dantcheva, "A video is worth more than 1000 lies. Comparing 3DCNN approaches for detecting deepfakes," 2020. [Online]. Available: https://apps.apple.com/gb/app/

[13] P. Majumdar, A. Agarwal, R. Singh, and M. Vatsa, "Evading Face Recognition via Partial Tampering of Faces," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2019, pp. 11–20. doi: 10.1109/CVPRW.2019.00008.

[14] H. Xu *et al.*, "Adversarial Attacks and Defenses in Images, Graphs and Text: A Review," Sep. 2019, [Online]. Available: http://arxiv.org/abs/1909.08072

[15] S. A. Aduwala, M. Arigala, S. Desai, H. J. Quan, and M. Eirinaki, "Deepfake Detection using GAN Discriminators," in *2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)*, Aug. 2021, pp. 69–77. doi: 10.1109/BigDataService52369.2021.00014.

[16] Y. Al-Dhabi and S. Zhang, "Deepfake Video Detection by Combining Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN)," in *2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE)*, Aug. 2021, pp. 236–241. doi: 10.1109/CSAIEE54046.2021.9543264.

[17] J. Hernandez-Ortega, R. Tolosana, J. Fierrez, and A. Morales, "DeepFakesON-Phys: DeepFakes Detection based on Heart Rate Estimation," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.00400

[18] K. Sun *et al.*, "Domain General Face Forgery Detection by Learning to Weight," 2021. [Online]. Available: https://github.com/skJack/LTW.

[19] A. Das, S. Das, and A. Dantcheva, "Demystifying Attention Mechanisms for Deepfake Detection," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, Dec. 2021, pp. 1–7. doi: 10.1109/FG52635.2021.9667026.

[20] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast Networks for Video Recognition," Dec. 2018.

[21] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time Face Capture and Reenactment of RGB Videos," Jul. 2020, [Online]. Available: http://arxiv.org/abs/2007.14808

[22] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," Oct. 2016, [Online]. Available: http://arxiv.org/abs/1610.02357

[23] J. Li *et al.*, "DSFD: Dual Shot Face Detector," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2019, pp. 5055–5064. doi: 10.1109/CVPR.2019.00520.

[24] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," May 2019.

[25] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection," Jan. 2020, [Online]. Available: http://arxiv.org/abs/2001.00179