DeepFake Detection using 3D-Xception Net with Discrete Fourier Transformation

Adeep Biswas
School of Computer Science and Engineering, Vellore Institute of Technology, Vellore
adeep.biswas2016@vitstudent.ac.in
Debayan Bhattacharya
School of Computer Science and Engineering, Vellore Institute of Technology, Vellore
debayan.bhattacharya2016@vitstudent.ac.in
Kakelli Anil Kumar*
Associate Professor, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore
anilsekumar@gmail.com

Received: 18/Aug/2020 Revised: 28/March/2021 Accepted: 01/June/2021

Abstract
The videos are more popular for sharing content on social media to capture the audience's attention. The artificial manipulation of videos is growing rapidly to make the videos flashy and interesting but they can easily misuse to spread false information on social media platforms. Deep Fake is a problematic method for the manipulation of videos in which artificial components are added to the video using emerging deep learning techniques. Due to the increase in the accuracy of deep fake generation methods, artificially created videos are no longer detectable and pose a major threat to social media users. To address this growing problem, we have proposed a new method for detecting deep fake videos using 3D Inflated Xception Net with Discrete Fourier Transformation. Xception Net was originally designed for application on 2D images only. The proposed method is the first attempt to use a 3D Xception Net for categorizing video-based data. The advantage of the proposed method is, it works on the whole video rather than the subset of frames while categorizing. Our proposed model was tested on the popular dataset Celeb-DF and achieved better accuracy.

Keywords: Computer Vision; DeepFake Detection; Xception Net; Video Manipulation.

1- Introduction

In recent times, the usage of videos has increased rapidly for different purposes such as marketing, news, and entertainment [1]. Along with the growing popularity of video-based content, some major social media platforms have come up as well and while these platforms have many benefits, they do not regulate or verify the videos being posted on their platform to a large extent [2]. Due to limited regulations and control systems, the users can easily manipulate videos artificially for malicious purposes such as spreading false political propaganda or causing disruptions in the financial market through spreading false rumours and fake information about military and research organizations by manipulating the satellite videos [3]. Due to such possibilities of malicious use, detecting such videos has become a serious issue [4]. One of the most common types of artificial videos is DeepFake. DeepFakes are videos or images which have been generated artificially using deep learning models [5]. DeepFakes rely on two types of artificial manipulation of videos mainly face swapping and facial re-enactment. Face swapping involves the replacement of a person's face in an image or a video with another face [6]. On the other hand, facial re-enactment is the process of creating the artificial head, lips, or any other facial feature movement to create a false narrative about a person's speech or expressions [7]. Both are artificial manipulations that are very harmful to social media users.

Deepfake generation technique is perfect and has some weaknesses that can be exploited to detect them. It is essential to determine various factors that can be used to differentiate a deepfake video from a genuine one. Some factors or features include the frequency of a blinking of eyelids in the video or even the anomalies in the head movement of a person [8]. The deepfake generation techniques have grown to be more sophisticated and accurate as well. The existing research work has focused to determine more robust detection techniques for deepfakes and accordingly, various newer techniques have been suggested such as the analysis of the colour hues in the video, the difference in the neural activation behaviour of the detection model, and the discrepancies in the convolutional traces of the video among other things [5].

* Corresponding Author

Although such newer detection techniques have achieved better accuracy results, it is essential to introduce advanced methods for better detection techniques continues. Due to the constantly evolving nature of deepfake generation, it is essential to introduce emerging detection techniques to overcome the challenges of the future [4]. The proposed work has proposed a new deepfake detection technique-3D Inflated Xception Net with Discrete Fourier Transformation which was able to achieve state-of-the-art accuracy results. To validate the performance of the proposed model, the publicly available deepfake benchmark dataset called Celeb-DF was used [9, 24]. Celeb-DF is a large-scale and high-quality deepfake dataset containing 5639 videos generated using various deep generation techniques [8, 9]. This data set is useful for the performance evaluation of various proposed algorithms/techniques for DeepFake detection and analysis. The main contribution of the method proposed work is it uses a 3D convolutional neural network model which takes the whole video as an input rather than extracting a subset of features from videos and using them as input parameters for categorization. Xception is proposed, and able to outperform most other pre-existing models in terms of accuracy as well as computational costs [10]. Despite having such promising results, its architecture has remained two-dimensional and relies on specific image frames extracted from a video for its categorization. The main drawback of such an approach is if the right video frame is not selected for the input, the video may get incorrectly categorized since artificial manipulations don't need to be done to all frames of the video when a deepfake is generated. This problem has been addressed in our proposed model by converting the 2-dimensional architecture of Xception net into 3-dimensions and initializing the network by pre-training it on static videos generated from a subset of images of the ImageNet dataset [25]. Furthermore, our model takes a two-stream approach and combines the results of the 3D Xception net with the results of a Discrete Fourier Transformation based classifier to account for all the parameters present in the spatial, temporal as well as frequency domains to achieve better accuracy results [26, 27].

The rest of the paper is arranged in the following manner- section 3 discusses the relevant research work which has already been done in this domain, section 4 illustrates details of the deepfake detection model proposed in this paper, section 5 specifies the configuration of our experimental setup and the results that were achieved by the proposed model and finally in section 6 we provide the conclusion and discuss the further scope for research work in this domain in the future.

2- Related Work

Significant breakthroughs have been made over the years now [28], there is still a lot of scope of improvements to handle the challenges in the domain of deepfake detection [29].

2-1- GAN Based Deepfake Generation

Generative Adversarial Networks (GAN) were introduced in 2014 and one of the most popular methods of generating deepfake videos [11, 30]. The main reason why GAN-based deepfake generation techniques have become important is they can adapt automatically and overcome any biases or weaknesses present in the network's generation process [12, 31].

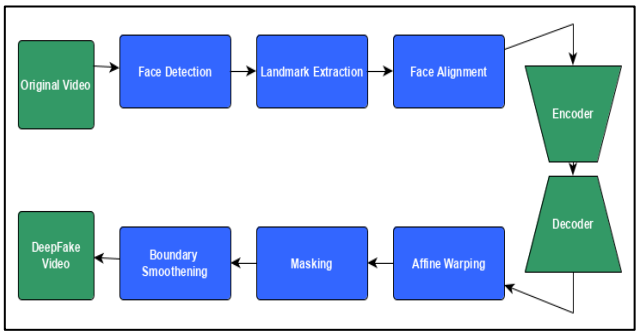

Fig. 1. DeepFake Synthesis Process

GAN has two major components, a generator, and a discriminator, both of which compete against each other for best results [32]. This works on any deepfake video outputted from the generator is passed to the discriminator which would try to determine whether the video inputted is fake or real and subsequently give feedback to the generator [13]. If the discriminator can classify the videos with high accuracy, it means that the generator function has some weaknesses that can be exploited. Based on this, the generator function would change its architecture and produce better deepfakes. On the other hand, if the discriminator is not able to make predictions accurately, then the generated deepfakes are of high quality and not easily distinguishable. Due to this constant process of feedback and improvement, GAN can overcome most detection methods eventually, causing those detection techniques to become obsolete [4]. Some such popular GAN methods proposed by researchers include AttGAN, StarGAN, StyleGAN, and PGGAN [11]. Due to the ability of GANs to change their generative architecture, traditional deepfake detection methods do not work on it and researchers are slowly moving towards a more forensic-based approach to counter GANs using the difference in color cues and pixel distribution [13]. The faces are targeted and detected in GAN Based Deepfake

Generation from the input video. The faces are aligned to the defined configuration using Landmark extraction and face alignment techniques as shown in figure 1. The encoder is used to detect the similarity of the facial expressions of the images or videos w. r.to target images. The GAN encoder and decoder are trained models, automated, the decoder can decode the target the facial expressions and also limit the re-construction errors. The targeted faces are wrapped, masking and boundary smoothing to the predefined configurations of original faces of the input images or video.

2-2- Existing DeepFake Detection Algorithms

The development of robust deepfake detection techniques has gained momentum in recent years due to their growing need. Some advanced deepfake detection methods have been presented to show the advancement and the variety of the methods. Li et al [14] proposed the detection of artificial manipulations in a video by locating the blending boundary between the combined real and fake potions of the video frames. Amerini et al [15] suggest a method of deepfake detection that relies on computing the optical flow of the given video. Fernando et al [16] put forth a technique of detecting deepfakes by utilizing memory networks to replicate human-like cognition of the social context of the video. Another interesting approach suggested by Venkatesh et al [17] is to detect a deepfake video by concentrating on the presence of morphing within its video frames. For this purpose, a form of context aggregation network is put forward. Jeon et al [18] have prioritized the computational cost of deepfake detection over its accuracy by proposing a lightweight neural network architecture that can utilize pre-existing and pre-trained classifier models. Zhang et al [19] has introduced a unique approach to tackling the deepfake problem by analyzing the difference in image compression ratios of the multiple video frames or images that are blend together through error level analysis. Dang et al [20] demonstrate a way of increasing the accuracy scores of deepfake detection classifiers by using eXtreme Gradient Boosting algorithm. Guera et al [35] has proposed a machine learning-based tool for automatic detection of manipulation traces in videos using CNN. Recurrent neural network (RNN) has used feature extraction and classification to find video manipulation. Afchar et al [36] has proposed an efficient automatic deep learning method for detect facial tampering in videos using Face-2-face and Deepfake. The work has achieved 95 to 98% accuracy using the above-mentioned methods. Sohrawardi et al [37] have proposed an efficient and robust system using artificial intelligence techniques for Deepkfake detection. Albahar et al [38] have analyzed the impact of the Deepfake in society hence they have introduced new techniques based on digital watermarking, facial detection techniques, and convolutional neural networks (CNNs). Using machine learning techniques, the method has achieved better accuracy in the detection of Deepfake videos. The proposed mechanism has achieved better results in terms of accuracy and efficiency in the detection of Deepfake. Therefore, from the review of these existing works, the problem of deepfake detection can be solved by using various techniques or parameters. The main challenge lies in determining these parameters, and other ways of improving the already existing detection technique.

This work deals with a modified version of Xception and the original Xception architecture proposed by Google researchers in 2017 [10]. Xception is a particular format of architecture or a particular way of arrangement of the different layers of activation present in a convolutional neural network model. This Xception model was created by modifying the previously benchmark CNN model called Inception and replacing its inception modules with depth-wise separable convolutions which made the network architecture more efficient and allowed a higher level of accuracy for the same number of input parameters. The main advantage of Xception over other deepfake detection techniques can achieve comparable and better accuracy results than the other existing detection techniques. The Xception is tested on popular datasets like FaceForenics [21] and DeeperForensics [22] through a 2-dimensional network in contrast to 3-dimensional or involving computationally expensive recurrent layers [33, 34]. Hence, Xception is an efficient algorithm and has the scope for better accuracy scores to make a perfect model for future purposes [37, 38].

3- Proposed Methodology

In this paper, a new algorithm for the detection of deepfake videos is proposed which involves 3D Inflated Xception Net with Discrete Fourier Transformation. The model takes in the whole video as an input and categorizes it in one of the two possible classifications namely fake or real based on the combined results of the 3D Xception stream and the 3D DFT stream. This allows the network to capture the overall spatial as well as frequency domain representations of the video from which local features. The local features are extracted subsequently in the convolutional layers and facilitating the classification while considering all parameters [39, 40]. The complete pipeline of the proposed algorithm including the data pre-processing and the 3D Xception and 3D DFT streams are presented in detail in the subsequent subsections.

3-1- Data Pre-Processing

The data pre-processing involved inputting the raw video into the 3D models to convert the videos into a uniform format with fixed dimensions for inputting into the model automatically one by one. It is necessary because the architecture of the 3D CNN used in our model needs inputs of fixed size. The raw video needed to be converted into a format that could be readable by the CNN and allow it to perform convolutional and padding operations. Initially, all video frames were extracted from the videos. Any video having lesser than 30 frames was discarded since they were too short. For the remaining videos, 30 frames were selected from the video. After that, each video frame is cropped into height and width dimensions of 299 by 299 pixels respectively. The cropped video frames are then converted into Numpy arrays and appended with each other to create 3-dimensional matrices of size 30 x 299 x 299 which could finally be inputted into the main classifier models.

3-2- Network Architecture

The model consists of two parallel streams- one consisting of the Xception model while the other one consisting of the Fourier transformation-based classifier. The overall structure is as shown in the figure 2.

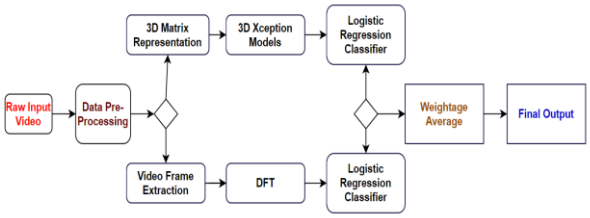

Fig. 2. Proposed Pipeline for DeepFake Detection

The Xception stream forms the first part of the pipeline which contains the 3D Xception model. A modified version of the Xception model is chosen because the original Xception architecture is a benchmark model for deepfake detection and outperforms other CNN architectural configurations for deepfake detection purposes. Furthermore, it showed as 2-dimensional, it showed great potential for being capable of being further enhanced by adding a third dimension to account for the temporal attributes of a video. The configuration of this model is the same as specified in the original Xception [2]. The only change is to the configuration instead of using 2-dimensional layers inside the CNN, they have been replaced with 3-dimensional convolutional layers by adding an extra dimension for the depth of the video to the already existing height and width dimensions. The newly constructed model was initialized randomly and pre-trained on inflated static videos formed from a subset of the ImageNet dataset. ImageNet dataset was chosen because it is one of the largest and most exhaustive datasets of images and has a demonstrated history of being a suitable option for pre-training models without imparting any biases to them. Furthermore, it needs to be noted that since ImageNet consists of only 2-dimensional images, these images had to first be converted into inflated videos by appending the same image 30 times to create a static video from the inflation of those given images. These inflated videos were then used to pre-train the Xception model to impart better weight initializations to the model than simple random initializations. The deepfake dataset called Celeb-DF was introduced to the model for training and testing purposes after completing pre-training of the model.

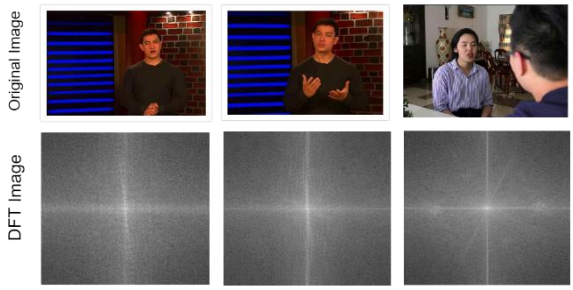

Fig. 3. Video Frames extracted from the Celeb-DF and their corresponding DFT Visualizations

The second stream consists of the Discrete Fourier Transformation module which is introduced to do a parallel frequency domain analysis on the video. The reason behind using a separate stream for frequency analysis of the videos is the most artificially generated videos can replicate the spatial properties of a video or an image. The replicating frequency and amplitude distribution of a genuine video is more difficult and often unaccounted the generation model and can be exploited to determine whether the video for artificially manipulated or not. For performing DFT transformation on the input videos, each of the video frames extracted from the raw videos during the data pre-processing phase is passed through a 2D discrete Fourier transformer and the output of these transformations is then appended to get an overall frequency domain representation. As shown in figure 3. Finally, this representation is compressed into a 2D matrix and then fed into a Logistic Regression algorithm-based classifier to get the output from the DFT stream. The logistic regression model is the best fit model as an exemplary classifier for binary and multi-class classification. The results of the proposed method have

shown that our hybrid classifier DFT with logistic regression model can differentiate the video samples of variable sizes with multiple features as real or fake. Once the output is given by the two streams, their outputs are combined to get the final result. The combination is done by taking a weighted average of the probability of the video being fake as predicted by the Xception stream and the DFT stream. Higher weightage is given to the output of the Xception model since it has a higher accuracy performance when used individually. The final probability is obtained from taking the weighted average of the individual probabilities of the two streams determines whether the video is fake or real by comparing it with a pre-determined threshold probability value.

4- Experimental Results

In this section, the experiments which are conducted to determine the performance of our proposed methodology are discussed in detail. This involves the overall experimental setup and resources used to achieve our results, the analysis of the outcome of the testing done on our proposed methodology as well as our model's performance comparison to other deepfake detection algorithms.

4-1- Experimental Setup

Due to the large size of video data as well as the high number of convolutional layers involved in the 3D Xception model, the training process of the prediction model was highly computationally expensive and hence, a 16GB GPU is used to run the code in our system. The rest of the configurations on which the code was run includes i7 processor, 8GB in-built RAM along an Ubuntu operating system environment. The complete implementation of the model is done on Python and Pytorch library is used for the implementation of the main 3D Xception module. The model is trained over 30 epochs to attain the final accuracy results. Besides the running environment, a publicly available benchmark dataset called Celeb-DF is used to test the performance of our proposed model. This particular dataset is chosen because it has a relatively large size containing 5639. Furthermore, the deepfake videos present in this dataset are of higher quality than those in previously available datasets and involve subjects of vast variations in terms of gender, age, and ethnicity, making this dataset suitable for replicating real-world deepfakes. A sample of this dataset is shown in Figure 4 where the original video and the deepfake were generated from it by performing artificial face-swapping.

Fig. 4. Real Video and the corresponding DeepFake generated

4-2- Results and Analysis

The performance of the model is measured using Area Under Curve (AUC) score of the Receiver Operating Characteristics graph since it is the standard parameter used for measuring the performance of the previous deepfake detection algorithms as well. Three different configurations of the proposed methodology are tested and their corresponding ROC curve and accuracy scores are presented in figure 5 and Table 1 respectively. The first configuration involves using the 3D Xception module directly without pre-training it on the ImageNet dataset or adding the DFT module to it. In the second case, the 3D Xception model is first pre-trained on ImageNet before training and testing it using the Celeb-DF dataset. As seen in the graph in Figure 5, the pre-training of the model significantly improves the performance of the model. Lastly, the pre-trained 3D Xception module is combined with the DFT module to further enhance the performance of the overall pipeline. Lastly, we compare the result achieved by the proposed methodology with the best AUC score of other prominent deepfake detection algorithms on the same Celeb-DF dataset [23]. From Table 2, our model's AUC score is among the best and comparable to some of the most advanced deepfake detection techniques. An important aspect here is that methods can perform 3D computations and those which are introduced to the Celeb-DF dataset during their training phase perform significantly better than the 2-dimensional algorithms.

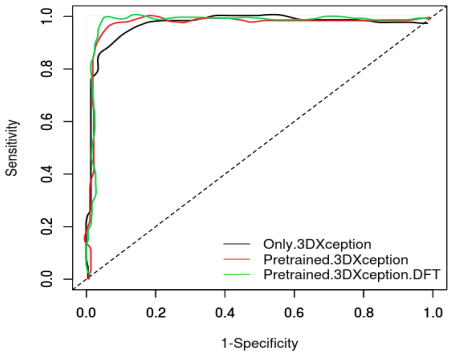

Fig. 5. ROC Curve for different configurations

Table 1. ROC-AUC % and Accuracy % of different configurations of the proposed methodology

Method	**ROC-AUC%**	**Accuracy %**
3D Xception	94.67	93.29
3D Xception (Pre-Trained with Image Net)	98.13	96.91
3D Xception- DFT (Pre- Trained with Image Net)	98.81	97.66

Table 2. ROC-AUC % Comparison of the Proposed Algorithm with other State-of-the-art DeepFake Detection Algorithms

Method	**Dimension**	**CALEB-DF Training**	**ROC-AUC %**
Two Stream	2D	NO	53.8
MESO4	2D	NO	54.8
MESOInception4	2D	NO	53.6
HEADPose	2D	NO	54.6
FWA	2D	NO	56.9
VA-MLP	2D	NO	55
VALogReg	2D	NO	55.1
Xception-raw	2D	NO	48.2
Xception-c23	2D	NO	65.3
Xception-c40	2D	NO	65.5
Mutli task	2D	NO	54.3
Capsule	2D	NO	57.5
DSP-FWA	2D	NO	64.6
DFT	3D	YES	66.8
X ception-Metric-Learning	3D	YES	99.2
RCN	3D	YES	74.87
R2Plus1D	3D	YES	99.43
I3D	3D	YES	97.59
MC3	3D	YES	99.3
R3D	3D	YES	99.73
3D Xception-DFT (The Proposed)	**3D**	**YES**	**98.81**

5- Conclusion and Future Scope

In this paper, we presented a new model for enhancing the performance of the pre-existing Xception algorithm. This was achieved by converting the whole architecture from being 2-dimensional into 3-dimensional space which is more suitable for handling the additional time-based dimension of videos. Furthermore, the paper also proposed a pipeline to combine the 3D Xception module with a frequency domain-based Fourier transformation model to achieve better results in terms of accuracy. Therefore, the model proposed in this paper accounts for all spatial, temporal as well as frequency domain-based parameters and hence can achieve results comparable to the existing state-of-the-art deepfake detection algorithms. The limitation of the proposed methodology is it increases the computational cost and complexity of the deepfake detection process. Thus, a common problem of the trade-off between efficiency and accuracy is created, each having its own merits for a particular use case. The future scope of this work would involve finding a way to make the proposed model more efficient while retaining the same level of accuracy. In real-time applications, speed is an important factor for the model to detect fake videos.

References

[1] Kumar, P., Vatsa, M., & Singh, R. Detecting face2face facial reenactment in videos. In The IEEE Winter Conference on Applications of Computer Vision, IEEE, 2020, pp. 2589-2597.

[2] Sabir, E., Cheng, J., Jaiswal, A., AbdAlmageed, W., Masi, I., & Natarajan, P. Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI), 2019, vol 3(1).

[3] Nguyen, T. T., Nguyen, C. M., Nguyen, D. T., Nguyen, D. T., & Nahavandi, S. Deep learning for deepfakes creation and detection, 2019, arXiv preprint arXiv:1909.11573.

[4] Lyu, S. Deepfake detection: Current challenges and next steps. In 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), IEEE, 2019, pp. 1-6.

[5] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. Deepfakes and beyond: A survey of face manipulation and fake detection. 2020, arXiv preprint arXiv:2001.00179.

[6] Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., & Nayar, S. K. Face swapping: automatically replacing faces in photographs. In ACM SIGGRAPH 2008 papers, 2008, pp. 1-8.

[7] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016 pp. 2387-2395.

[8] Tolosana, R., Romero-Tapiador, S., Fierrez, J., & Vera-Rodriguez, R. DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance, 2020, arXiv preprint arXiv:2004.07532.

[9] Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. Celeb-df: A new dataset for deepfake forensics, 2019, arXiv preprint arXiv:1909.12962.

[10] Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251-1258.

[11] Huang, Y., Juefei-Xu, F., Wang, R., Xie, X., Ma, L., Li, J., ... & Pu, G. FakeLocator: Robust Localization of GAN-Based Face Manipulations via Semantic Segmentation Networks with Bells and Whistles, 2020, arXiv preprint arXiv:2001.09598.

[12] Nirkin, Y., Keller, Y., & Hassner, T. FSGAN: Subject agnostic face swapping and reenactment. In Proceedings of the IEEE international conference on computer vision, 2019, pp. 7184-7193.

[13] McCloskey, S., & Albright, M. Detecting gan-generated imagery using color cues, 2018, arXiv preprint arXiv:1812.08247.

[14] Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. Face x-ray for more general face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5001-5010.

[15] Amerini, I., Galteri, L., Caldelli, R., & Del Bimbo, A. Deepfake video detection through optical flow based cnn. In Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019.

[16] Fernando, T., Fookes, C., Denman, S., & Sridharan, S. Exploiting human social cognition for the detection of fake and fraudulent faces via memory networks. 2019, arXiv preprint arXiv:1911.07844.

[17] Venkatesh, S., Ramachandra, R., Raja, K., Spreeuwers, L., Veldhuis, R., & Busch, C. Detecting morphed face attacks using residual noise from deep multi-scale context aggregation network. In The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 280-289.

[18] Jeon, H., Bang, Y., & Woo, S. S. FDFtNet: Facing Off Fake Images using Fake Detection Fine-tuning Network, 2020, arXiv preprint arXiv:2001.01265.

[19] Zhang, W., Zhao, C., & Li, Y. A Novel Counterfeit Feature Extraction Technique for Exposing Face-Swap Images Based on Deep Learning and Error Level Analysis. Entropy, 2020, vol 22(2), no. 249. ##

[20] Dang, L. M., Min, K., Lee, S., Han, D., & Moon, H. Tampered and computer-generated face images identification based on deep learning. Applied Sciences, 2020, vol 10(2), no. 505.

[21] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. Faceforensics: A large-scale video dataset for forgery detection in human faces, 2018, arXiv preprint arXiv:1803.09179.

[22] Jiang, L., Li, R., Wu, W., Qian, C., & Loy, C. C. DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2889-2898.

[23] de Lima, O., Franklin, S., Basu, S., Karwoski, B., & George, A. Deepfake Detection using Spatiotemporal Convolutional Networks, 2020, arXiv preprint arXiv:2006.14749.

[24] Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3207-3216.

[25] Chollet, F. Xception: deep learning with depthwise separable convolutions, in: 2017 IEEE conference on computer vision and pattern recognition CVPR, 2017.

[26] Kaiser, L., Gomez, A. N., & Chollet, F. Depthwise separable convolutions for neural machine translation, 2017, arXiv preprint arXiv:1706.03059.

[27] Rahimian, E., Zabihi, S., Atashzar, S. F., Asif, A., & Mohammadi, A. XceptionTime: Independent Time-Window Xceptiontime Architecture for Hand Gesture Classification. In ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 1304-1308.

[28] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. Deepfakes and beyond: A survey of face manipulation and fake detection, 2020, arXiv preprint arXiv:2001.00179.

[29] Guarnera, L., Giudice, O., & Battiato, S. DeepFake Detection by Analyzing Convolutional Traces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 666-667.

[30] Zhang, J., Salehizadeh, M., & Diller, E. Parallel pick and place using two independent untethered mobile magnetic microgrippers in IEEE International Conference on Robotics and Automation, 2018.

[31] Bau, D., Zhu, J. Y., Strobelt, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., & Torralba, A. Visualizing and understanding generative adversarial networks, 2019, arXiv preprint arXiv:1901.09887.

[32] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. Generative adversarial networks: An overview. IEEE Signal Processing Magazine, vol 35(1), 2018, pp 53-65.

[33] Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. Deepfakes: Trick or treat?. Business Horizons, 2020, vol 63(2), pp 135-146.
[34] Wang, J., Liu, A., & Xiao, J. Video-Based Pig Recognition with Feature-Integrated Transfer Learning. In Chinese Conference on Biometric Recognition, Springer, Cham, 2018, pp 620-631.
[35] Güera, D., & Delp, E. J. 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.
[36] 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 (WIFS), IEEE, 2018, pp. 1-7.
[37] Sohrawardi, S. J., Chintha, A., Thai, B., Seng, S., Hickerson, A., Ptucha, R., & Wright, M. Poster: Towards robust open-world detection of deepfakes. In Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, 2019, pp. 2613-2615.
[38] Albahar, M., & Almalki, J. Deepfakes: Threats and countermeasures systematic review. Journal of Theoretical and Applied Information Technology, vol 97(22), 2019, pp 3242-3250.
[39] Maksutov, A. A., Morozov, V. O., Lavrenov, A. A., & Smirnov, A. S. Methods of Deepfake Detection Based on Machine Learning. In 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), IEEE, 2020, pp. 408-411.
[40] Korshunov, P., & Marcel, S. Deepfakes: a new threat to face recognition assessment and detection, 2018, arXiv preprint arXiv:1812.08685.

Adeep Biswas is a graduate in B. Tech in Computer Science and Engineering from Vellore Institute of Technology. He looks forward to pursuing his Graduate Degree in information and communication technologies. His research interests include Image Processing, Information retrieval, Recommendation systems, Network Security, Web development.

Debayan Bhattacharya is a graduate in B. Tech in Computer Science and Engineering from Vellore Institute of Technology. He looks forward to pursuing his Graduate Degree in information and technology. His research interests include Image Processing, Network Security, Web development and computer vision.

Kakelli Anil Kumar is an Associate Professor of the School of Computer Science and Engineering at the Vellore Institute of Technology (VIT), Vellore, TN, India. He earned his Ph.D. in Computer Science and Engineering from Jawaharlal Nehru Technological University (JNTUH) Hyderabad in 2017, and graduated in 2009 and under-graduated in 2003 from the same university. He started his teaching career in 2004 and worked as an Assistant Professor, and Associate Professor and HOD in various reputed institutions of India. His current research includes wireless sensor networks, internet of things (IoT), cyber security and digital forensics, Malware analysis, block-chain and crypto-currency. He has published over 40 research articles in reputed peer reviewed international journals and conferences.