Phase I Report
On

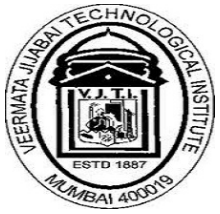
"Detection Of Deepfake videos by Using Combination of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN)"

Submitted
in partial fulfilment of the requirement for
the degree of Master of Technology
in Computer Engineering

by

Vinay Deshmukh
(192050008)

Under the guidance of
Prof. Shraddha S. Suratkar



DEPARTMENT OF COMPUTER ENGINEERING
VEERMATA JIJABAI TECHNOLOGICAL INSTITUTE
(An Autonomous Institute Affiliated to Mumbai University)
(Central Technological Institute, Maharashtra State)
Matunga, MUMBAI - 400019

A.Y. 2020-2021

STATEMENT OF CANDIDATE

I state that work embodied in this Project entitled **"Detection Of Deepfake videos by Using Combination of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN)"** form my own contribution of work under the guidance of **Prof. Shraddha S. Suratkar** at the Department of Computer Engineering, Veermata Jijabai Technological Institute, Mumbai. The report reflects the work done during the period of candidature but may include related preliminary material provided that it has not contributed to an award of previous degree. No part of this work has been used by us for the requirement of another degree except where explicitly stated in the body of the text and the attached statement.

Vinay Deshmukh
Roll No. : 192050008
Date :
Place : VJTI, Mumbai

APPROVAL SHEET

The stage I report on **"Detection Of Deepfake videos by Using Combination of Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN)"** by Vinay Deshmukh is found to be satisfactory and is approved for the Degree of Master of Technology.

Prof. S. S. Suratkar
M.Tech Supervisor

Examiner

Examiner

Examiner

Place : Veermata Jijabai Technological Institute, Mumbai
Date :

Acknowledgements

I would like to thank all those people whose support and cooperation has been an invaluable asset during the course of this Project. I would also like to thank our Guide **Prof. S. S. Suratkar** for guiding me throughout this project and giving it the present shape. With his encouraging words and reminders, he empowered me to improve in personal as well as academic life. It would have been impossible to complete the project without his support, valuable suggestions, criticism, encouragement and guidance.

I convey my gratitude also to **Dr. M. M. Chandane**, Head of Department for his motivation and providing various facilities, which helped us greatly in the whole process of this project.

I am also grateful to all other teaching and non-teaching staff members of the Computer Engineering and Information Technology Department for directly or indirectly helping us for the completion of project and the resources provided.

Vinay Deshmukh

Contents

List of Figures

Chapter 1

Synopsis

1.1 Abstract

Nowadays, people are facing an emerging problem called deepfake videos, these video are created using deep learning technology. These videos are created to cause threats to privacy, reputation and so on. Sometimes deepfake videos created using latest algorithms could be hard to distinguish with naked eye. That's why we need better algorithms to detect deepfake.

The system we are going to present is based on combination CNN and RNN, as research shows that using CNN and RNN combined achieves better results.We are going to use pre-trained CNN model called EfficientNet, using this we save the time of training the model from scratch. The proposed system uses CNN to extract features and these extracted features are used to train the LSTM network. Using CNN and RNN combined we capture the inter frames as well as intra frames features which will be used to detect if the video is real or fake.

Keywords: Deep learning, DeepFake detection, Neural network

Chapter 2

Introduction

2.1 Deepfake

"Deepfake" word is the combination of "Deep learning" and "fake", these are the AI generated videos in which a person in a video or image is replaced with someone else. The idea of swapping faces on photo is not new, we can find examples of such photos made in 19th century, back then these photos were made by hand. However, when the idea of neural networks became popular and with advancement in computational field, people began to use this technology to create such deepfake videos and images. Nowadays we can download and run such programs which can swap our face with others. Today, none of us will be surprised by apps like FaceApp, Snapchat that have an ability to swap faces with good quality and make us funny.

In January 2018, a desktop application called FakeApp was launched, This app allows users to easily create and share videos with their faces swapped with each other, after this app many such apps were launched like FaceSwap, DeepFaceLab. Larger companies also started to use deepfakes. The Japanese AI company DataGrid made a full body deepfake that can create a person from scratch, they intend to use these for fashion and apparel. A mobile deepfake app, Impressions, was launched in March 2020. It was the first app for the creation of celebrity deepfake videos from mobile phones. Today we can find various desktop as well as mobile apps which we can use to create a good quality deepfake.

2.2 Deepfake Generation Techniques

The mechanism for deepfake creation is deep learning models such as autoencoders and generative adversarial networks(GAN), which have been applied widely in the computer vision domain

2.2.1 Autoencoders

It is well known neural network to create deepfake videos. Autoencoders contains two parts encoder and decoder. Function of encoder is the compression of input. First the input is represented in smaller latent space using encoder. The function of decoder is opposite of encoder to get original data from compressed data. By using encoder of one image with the decoder of other we can create a deepfake. For deepfake creation this encoder and decoder only applied on facial region. Figure 2.1 shows the architecture of autoencoder which is used to create deepfake image. As shown in figure encoder of face A is used with the decoder of face B to get reconstructed face B from A. Some example which uses these techniques to create deepfae are Faceswap, DFaker, DeepFakeLab and etc.

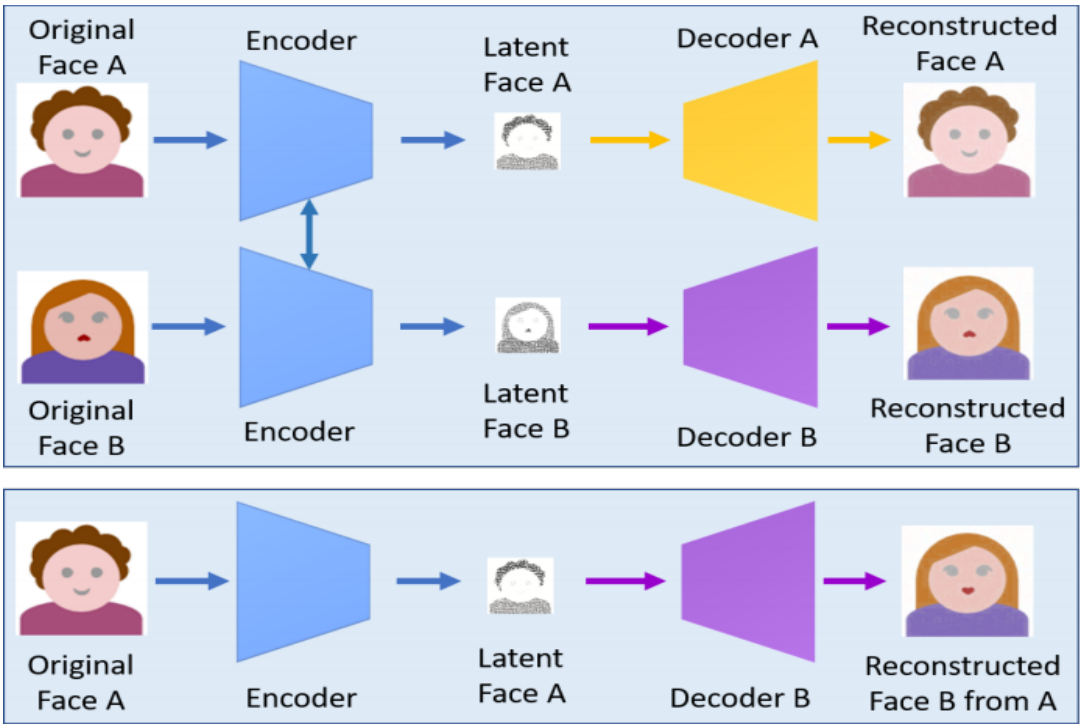


Figure 2.1: Autoencoder Technique

2.2.2 GAN

GAN is the advancement in the autoencoder technique to generate deepfake. In GAN there are two differnet deep neural networks, generator and discriminator. Generator in GAN works similar to autoencoder, but we can achieve better results because discriminator net is rejecting some bad examples. Job of discriminator is to reject bad deepfake example and generator keeps producing the deepfake until it successfully fools the discriminator, that makes such fakes more similar to real videos and also makes them harder to recognize by naked eyes. Some of the open-source

projects are using this technique for example, Faceswap-GAN. Figure 2.2 shows the architecture of GAN network.

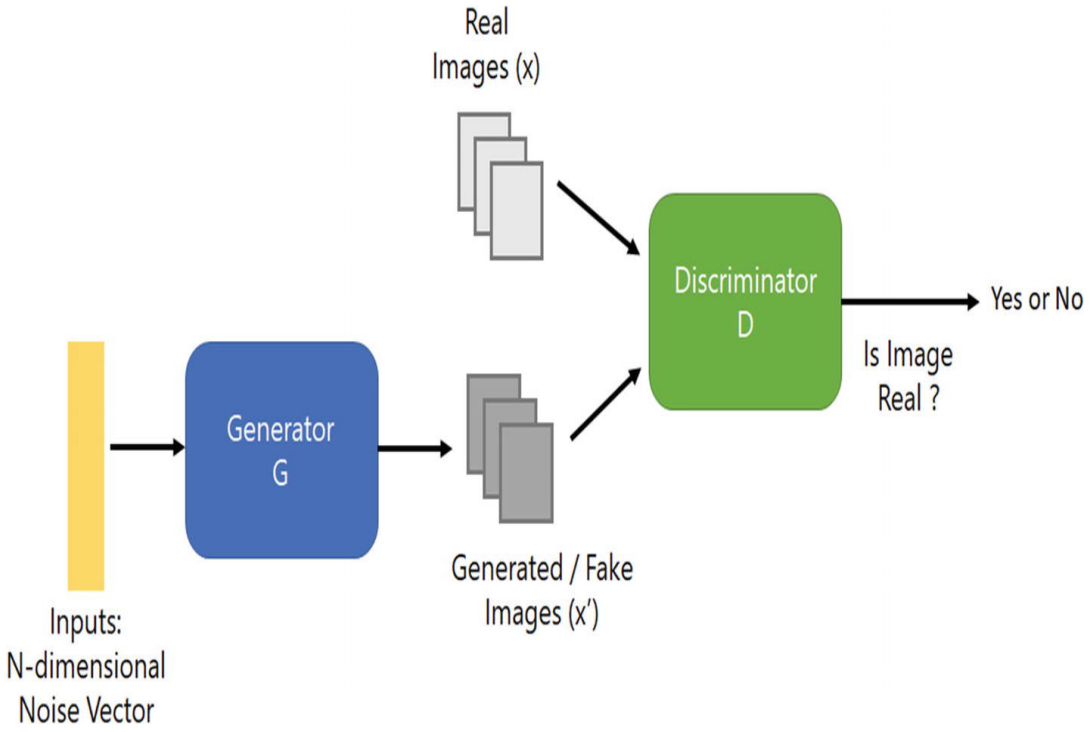


Figure 2.2: GAN Technique

2.3 Motivation

Now days everyone has access to the deepfake generation applications. At first these techniques were mainly used to create funny videos of celebrities and politicians, making them say funny things. However, it would be just as easy to create a deepfake of an higher authority person delivering fake speech, it can easily used to cause political or religion tension between nations also or to fool the public by dropping the video of politician just before the elections , it is also used for Putting celebrities face on porn videos harming their reputation or to create chaos in financial markets by creating fake news. The first deepfake video emerged in 2017 where face of a celebrity was swapped to face of a porn actor since then deepfake generation techniques are getting more advance. With the advancement in deepfake we should also improve the techniques used to detect deepfake.

2.4 Problem Statement

Previously proposed systems which uses CNN and RNN combined managed to achieve the average accuracy above 90% but the test were conducted on same distri-

bution where the CNN and RNN are trained, when such techniques will be applied on different distribution, it will fail to achieve same accuracy. Apart from hybrid networks which uses CNN and RNN combined other techniques fail to achieve accuracy up to the mark. For increasing the accuracy scaling the network in terms of depth, width or high quality input is an option but scaling the network not always actually increase the accuracy, the scaling should be uniform for better results.

2.5 Objectives

1. To create a deep neural network which can identify deepfake videos accurately, by using transfer learning approach.
2. To select pre-pertained network which will be used to maximize the accuracy of the network.
3. To implement CNN to extract frame level features.
4. To implement LSTM which uses extracted features for training and produce the final output.
5. Create the dataset which is combination of different distributions, because by using videos from different distribution the results will be consistent.

Chapter 3

Literature Survey

3.1 Use of CNN and RNN Based Hybrid Model for Deepfake Detection

In paper [1] proposed a CNN and LSTM based deepfake detection method. This method uses convolutional Neural Network (CNN) and stores the abnormal features for training. A total of 512 facial landmarks were extracted and compared. Parameters such as eye-blinking lip-synch, eyebrows movement, and position, are few main deciding factors that classify into real or counterfeit visual data. Instead of storing frames, main features extracted from CNN are stored in LSTM using mean pooling. The Recurrent Neural Network (RNN) pipeline learns based on these features-fed inputs from CNN and then evaluates results. The model was trained with the network of videos consisting of their real and fake, collected from multiple websites. This training compares real video with its deepfake video and CNN extract the abnormal features only. The proposed algorithm and designed network set a new benchmark for detecting the visual counterfeits and show how this system can achieve competitive results on any fake generated video or image. Problem with this method is that without real video first it is challenging to extract abnormal features.

The Methods which uses CNN and RNN combined together will use CNN for feature extraction and RNN is used to produce final output. One of such method is presented in [5] the CNN feature extractor used in this method is InceptionV3 and RNN used is LSTM with 0.5 chance of dropout, the accuracy achieved is around 99% but this accuracy achieved is only on 600 video data-set that they have created. Another method which uses CNN and RNN combined is proposed in [14] this method uses Eulerian Motion Magnification along with InceptionV3 and LSTM. Eulerian Motion magnification is used to magnify the facial region of video and then CNN is used to extract inter frame features and LSTM used to extract intra frame

features. The accuracy achieved on FaceForensic++ data set is 99.25%.

3.2 Deepfake Detection Using CNN Based Classifier

Some proposed method uses CNN classifier with other algorithms to proudce the final output. In [11] proposed a architecture which uses ResNet classifier along with inconsistent head pose estimator and fast super resolution CNN model. This method achieves accuracy of 95.5% but ResNet classifier alone can achieve accuracy of 94.9% on UADFV dataset. Method proposed in [9] optical flow fields are used with CNN. Optical flow is a vector field which is computed on two consecutive frames to extract apparent motion between the observer and the scene itself.Using this optical flow vector and VGG16 accuracy achieved is 81% when used with ResNet50 accuracy achieved is 75%. This accuracy achieved is achieved on only FaceForensic++ dataset. In [8] proposed a method to detect the deepfake using discriminator used in GAN network. The proposed method uses two discriminators to increase the capability of deepfake detection. The accuracy achieved with this method is less but the results were consistent. In [3] proposed a method which uses CNN to capture distinctive artifacts in deepfake video. The authors of this paper used image processing for creation of deepfake videos. This method is tested on VGG16, ResNet50, ResNet101, ResNet152 among these ResNet50 outperforms others with accuracy 97.4%. In [16] proposed a method which uses EfficientNetB5 with MTCNN the accuracy achieved is 92.61% on DFDC dataset. In [4] proposed two deep neural network Meso4 and MesoInception4 both have very number of layers still manages to achieve average classification score of 0.89 for Meso4 and 0.971 for MesoInception4. This method mainly focused on two Deepfake video generation techniques Deepfake and Face2Face, so while testing the dataset used was generated using these two methods only.

3.3 Other Methods of Deepfake Detection

There are also methods in deepfake detection which does not include any type of deep neural network. In [15] proposed a method which compares the blur and sharpness of facial region with the background, and based on that it detects the deepfake video. The accuracy achieved with this method on UADFV dataset is 90%. In [13] provides a architecture which uses less number of frames per video to assess its realism. this method uses facenet with the metric learning approach using a triplet net-

work architecture. In [12] proposed a DeepVision algorithm which uses frequency of eye blinking in video for detection of deepfake. Various factors which effect the blinking pattern (age, time of day, gender, emotional state etc.) using these factors as a input the proposed algorithm can predict if the video is fake or not. For this to work efficiently the input should be precise. In [10] proposed a method which uses SVM classifier with edge features detection algorithms such as HOG, SURF, KAZE, etc. Using these SVM is trained for classification. Using HOG the accuracy achieved is maximum i.e. 94%. In [6] proposed a method which uses head position for detection of deepfake. The proposed method first extract 2D facial landmark to create 3D model of head and deepfake videos show inconsistency in 3D head poses, that inconsistency is captured using SVM classifier. These methods which does not use any deep neural network does not achieve the acurracy up to the mark.

3.4 EfficientNet: Model based on Uniform Scaling

Convolutional Neural Networks (ConvNets) are commonly developed at a fixed resource budget, and then scaled up for better accuracy if more resources are available. Commonly scaling is done on depth, width, and image size. In paper [18] proposed a new scaling method that uniformly scales all dimensions of depth,width and resolution using a simple yet highly effective compound coefficient. The compound scaling is based on fact that different scaling dimensions are not independent. By using this compound coefficient method on ResNet and MobileNets this paper also proposed a new model architecture called EfficientNet. Which achieves 84.3% top-1 accuracy on ImageNet, while being 8.4x smaller and 6.1x faster on inference than the best existing ConvNet Gpipe, while Inceptionv4 achieves 80.1% top-1 accuracy and ResNet50 achieves 76.0% top-1 accuracy.

3.5 Literature Gap

1. Previously proposed models shows that we can achieve good accuracy by using CNN and RNN combined but the results were only based on testing the model on same distribution or testing model on very few examples.
2. Most of the proposed model which does not use deep neural network fail to achieve accuracy up to the mark.
3. Combining more methods of deepfake detection together not always achieves better results.

4. For models which uses CNN and RNN combined, better network can be selected for feature extraction to achieve better results.

Chapter 4

System Design

4.1 Architecture

Proposed system will make use of CNN and RNN combined. The CNN used is EfficientNet for feature extraction and LSTM is trained based on those features.

The architecture of the proposed system is shown in figure 4.1.

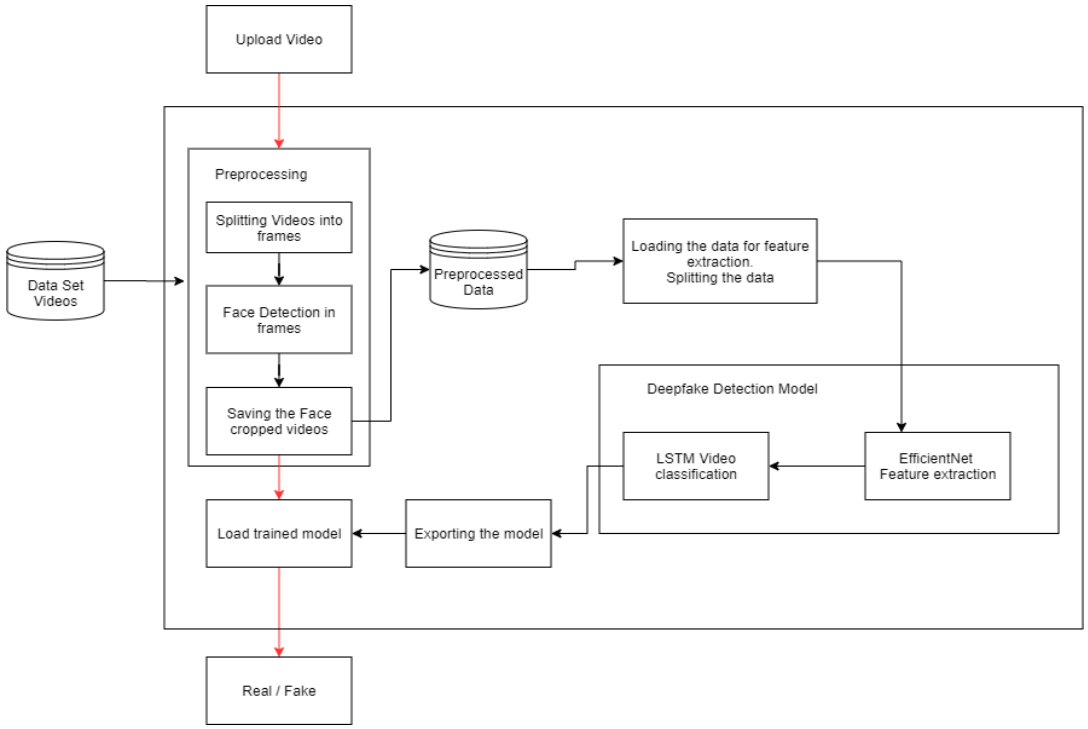


Figure 4.1: System Architecture

4.2 Steps

1. First we have to collect the data set of real and fake videos it should be labeled data set, while collecting data set we have to make sure that the videos come

from different distribution for better results.

2. After collecting the data set we have to do pre-processing of data i.e. we have to extract the frames from the videos and then extract faces region from the frames.
3. We have to split the data into training set and testing set, 70% for training and 30% for testing.
4. After extracting faces from the frames we have to pass that data to pre-trained EfficientNet for feature extraction.
5. The output feature vector of CNN is passed to LSTM for training purpose, after training LSTM will produce the final output to check the accuracy we will use testing set.

Chapter 5

Project Planning

5.1 Work plan

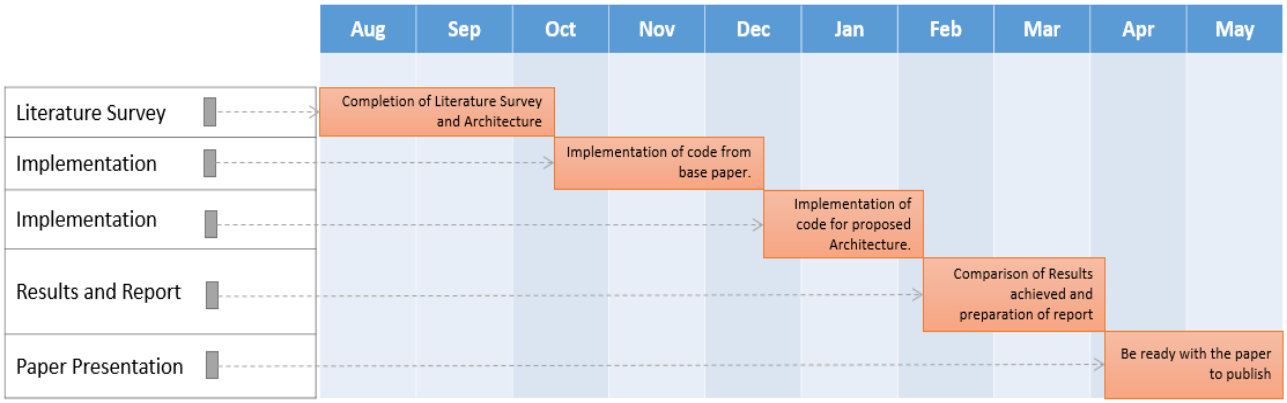


Figure 5.1: Work plan

Chapter 6

Conclusion

6.1 Conclusion

With the advancement in deepfake video generation techniques we should also try to improve the detection techniques which are accurate as well as efficient. The methods of deepfake detection which does not include any deep neural network are deepfake detection using inconsistent head poses[6], human eye blinking pattern[12], but these methods also depend on various other factor which makes them less efficient and also the accuracy of such methods is less compared to methods which uses deep neural network.

6.2 Future Work

This proposed architecture uses EfficientNet[18], this CNN architecture is better than previously used techniques such as Inceptionv3, ResNet and MobileNets for feature extraction. EfficientNet is smaller in size and also achieves better top-1% results as compared to all other networks. EfficientNet will be used for feature extraction. Using this EfficientNet with LSTM we can achieve better result than previous methods.

References

[1] M. F. Hashmi, B. K. K. Ashish, A. G. Keskar, N. D. Bokde, J. H. Yoon and Z. W. Geem, "An Exploratory Analysis on Visual Counterfeits Using Conv-LSTM Hybrid Architecture," in IEEE Access, vol. 8, pp. 101293-101308, 2020, doi: 10.1109/ACCESS.2020.2998330.

[2] H. Li, Y. Huang and Z. Zhang, "An Improved Faster R-CNN for Same Object Retrieval," in IEEE Access, vol. 5, pp. 13665-13676, 2017, doi: 10.1109/ACCESS.2017.2729943.

[3] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. arXiv preprint arXiv:1811.00656, 2018

[4] D. Afchar, V. Nozick, J. Yamagishi and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, Hong Kong, 2018, pp. 1-7, doi: 10.1109/WIFS.2018.8630761.

[5] D. Güera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6, doi: 10.1109/AVSS.2018.8639163.

[6] X. Yang, Y. Li and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 8261-8265, doi: 10.1109/ICASSP.2019.8683164.

[7] P. Korshunov and S. Marcel, "Vulnerability assessment and detection of Deepfake videos," 2019 International Conference on Biometrics (ICB), Crete, Greece, 2019, pp. 1-6, doi: 10.1109/ICB45273.2019.8987375.

[8] J. Baek, Y. Yoo and S. Bae, "Generative Adversarial Ensemble Learning for Face Forensics," in IEEE Access, vol. 8, pp. 45421-45431, 2020, doi: 10.1109/ACCESS.2020.2968612.

[9] I. Amerini, L. Galteri, R. Caldelli and A. Del Bimbo, "Deepfake Video Detection through Optical Flow Based CNN," 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South), 2019, pp. 1205-1207, doi: 10.1109/ICCVW.2019.00152.

[10] F. F. Kharbat, T. Elamsy, A. Mahmoud and R. Abdullah, "Image Feature Detectors for Deepfake Video Detection," 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, 2019, pp. 1-4, doi: 10.1109/AICCSA47632.2019.9035360.

[11] N. S. Ivanov, A. V. Arzhskov and V. G. Ivanenko, "Combining Deep Learning and Super-Resolution Algorithms for Deep Fake Detection," 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), St. Petersburg and Moscow, Russia, 2020, pp. 326-328, doi: 10.1109/EIConRus49466.2020.9039498.

[12] T. Jung, S. Kim and K. Kim, "DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern," in IEEE Access, vol. 8, pp. 83144-83154, 2020, doi: 10.1109/ACCESS.2020.2988660.

[13] M. F. Hashmi, B. K. K. Ashish, A. G. Keskar, NA. Kumar, A. Bhavsar and R. Verma, "Detecting Deepfakes with Metric Learning," 2020 8th International Workshop on Biometrics and Forensics (IWBF), Porto, Portugal, 2020, pp. 1-6, doi: 10.1109/IWBF49977.2020.9107962.

[14] Fei, J., Xia, Z., Yu, P. et al. "Exposing AI-generated videos with motion magnification." Multimed Tools Appl (2020). https://doi.org/10.1007/s11042-020-09147-3

[15] M. A. Younus and T. M. Hasan, "Effective and Fast DeepFake Detection Method Based on Haar Wavelet Transform," 2020 International Conference on Computer Science and Software Engineering (CSASE), Duhok, Iraq, 2020, pp. 186-190, doi: 10.1109/CSASE48920.2020.9142077.

[16] AD. M. Montserrat et al., "Deepfakes Detection with Automatic Face Weighting," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2020, pp. 2851-2859, doi: 10.1109/CVPRW50498.2020.00342.

[17] J. Deng, W. Dong, R. Socher, L. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Com-

puter Vision and Pattern Recognition, Miami, FL, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[18] Mingxing Tan, Quoc V. Le "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks" (2019) Cornell University https://arxiv.org/abs/1905.11946