Deepfake Detection: A Modern Survey on Methods and Techniques

Ravi Maithrey Regulagedda
School of Computer Science and Engineering, VIT University
Tiruvalam Road, Katpadi, Vellore, Tamil Nadu, 632014
ravi.maithrey2018@vitstudent.ac.in

Potula Sri Rupin
School of Computer Science and Engineering, VIT University,
Tiruvalam Road, Katpadi, Vellore, Tamil Nādu, 632014
srirupin.potula2019@vitstudent.ac.in

Chinta Aaron John
School of Computer Science and Engineering, VIT University,
Tiruvalam Road, Katpadi, Vellore, Tamil Nadu, 632014
chintaaaron.john2018@vitstudent.ac.in

Abstract- Deepfakes are images or videos that have been modified to changes their faces or other parts of the image to change the subject. Usually, the faces of one person are replaced by another. This deepfake generation is done primarily via Generative Adversarial Networks (GANs). This makes the deepfake have certain characteristics that are unique to the generation process through which they can be detected. This type of modification has a great capacity for producing fraudulent media which can be used to cause confusion and harm. Therefore, a method for the detection of deepfakes is of the utmost urgency. As deepfake technology becomes more and more advanced, detection methods need to keep up. In this paper, we present a survey and review of the most recent deepfake detection methods in order to facilitate any further research in the domain. Each technique has been carefully described and their accuracy on different common deepfake datasets has been described. This should provide a starting point for anyone wishing to perform further research on any of these techniques or come up with one on their own.

Keywords- Deepfake, Deepfake Detection, Generative Adversarial Networks, Deep Learning, Unsupervised Learning

I. INTRODUCTION

Goodfellow et al introduced the most powerful networks which is Generative Adversarial Networks in the year 2014[1]. The meaning of the generative is to create new images based on the training set given. The main usage of GANs has been to generate images. GANs aim to solve the generative problem where the task is to learn the underlying distribution of data and then to use this to generate data that is similar to the distribution it has learned from. In this usage it has many mathematical applications. However, in daily usage, as described above, GANs have been used to learn data distributions to reproduce data as if from the same original source. Beyond just image generation, this also finds use in dataset augmentation, which is of great importance in training better machine learning models.

GANs use a supervised learning technique where the entire model is made up of two neural networks which feed into each other. This is a framework that consists of a generator and a discriminator. As one generates fakes (generator) and the other identifies them (discriminator) [2]. The idea behind is that one network supervises the other network. Thus, it is a combination of unsupervised and supervised techniques. The GAN is said to have learned from its training data when the discriminator can no longer distinguish that the images and any other produced by the generator are fake.

GANs have also recently found usage in the generation and dissemination of a type of images known as deepfakes. These deepfakes pose a great problem in terms of their impact. The current widespread prevalence and use of deepfake generation methods is a problem which of both technological and social importance. The detection of deepfakes which have been generated by such GANs is the focus of this paper. Here we will present current advances in deepfake detection methods along with their mathematical basis and accuracy on a few common datasets. This should provide an idea of the current advances this field. The methods chosen here were done so primarily on the basis of how recently they were published as well their accuracy scores. We hope this survey provides useful information on Deepfake detection techniques.

II. DEEPFAKE AND THEIR IMPACT

In simple terms, swapping the faces of persons from any source such as videos or images by the use of machine learning and image generation techniques is known as Deepfake [3]. Because deepfake technology inherently assumes the identity of a person this causes adverse effects to the privacy and individuality of a person [4]. The most common type of deepfake, and the one that is most widespread in terms of use, is the face swap. Some deepfakes are just static images, while others are full videos where each frame in the video is deepfaked.

Generative Adversarial Networks are used to generate Deepfake. As discussed above this results in improper usage of images and videos. The prevalence of deepfake generation technology has grown to such an extent that deepfakes can be generated on via applications on mobile phones or on the web, with minimal resources to work with.[5] One such app is FACEAPP, where any person can transform the faces, aging, etc. where a person with minimum experience or knowledge can do this. This is a huge problem as it can cause the generation and wrongful use of such deepfakes very easily

Besides this, another problem arises due to the generation of fake news by deepfake. Fake news is a huge problem these days, causing political and real-world problems due to the false information it disseminates. Deepfakes make it easier to spread fake news as people are more likely to believe a video or an image than just a headline and text. Faces-swapping of celebrities, officials, presidents and others degrades their reputation in the society. Most of the videos were forged for the sake of entertainment on late night TV [6], but there are cases where more nefarious purposes were in mind, ranging from causing a loss of confidence in governmental leaders to sowing discord and confusion by sending harmful messages which have been deepfaked to grant legitimacy.

Deepfake causes threat to judicial systems as evidence can be tampered. Soon, it will be hard to produce images or videos as evidence in court without proving that they are not deepfakes, which is another issue to be tackled. Tampering the images of politicians, celebrities and others and releasing fake videos on social media sites results in loss of reputation of these people, beyond causing the spread of fake news which is a problem in its own. So, to lower the prevalence of these problems deepfake detection is of the utmost importance.

Due to the reasons mentioned above, detection of deep fakes takes on a very important role. In the past few years great strides have been made in this avenue. This paper takes a look at these methods in order to provide a starting point for any research being made in this avenue.

III. DEEPFAKE DETECTION

Below we present some of the latest developments in the field of deepfake detection, giving a detailed description of each method, the standard it has been tested against and results obtained in order to facilitate understanding. First a basic description of the background of the method is provided. Then, a mathematical description of the detection method is described to provide insight into the working behind this detection method. Finally, the accuracy scores of each of these detection methods is shown in order to facilitate easy comparison.

3.1 Deepfake Detection using Convolutional Traces

Exposing the convolutional traces on images is a new approach to fight Deepfake which was developed by [7]. In this paper, a new approach was developed that is the extraction of Convolutional Traces that are formed during the generation phase of running the GAN during deepfake creation. Detection of Deepfake is difficult for humans but results show that they can be easily detected by Convolutional Neural Networks (CNNs) [8] but there are few limitations such as explainability and capability to generalize.

To overcome this, extraction of the fingerprints that were left behind GANs is done using Expectation Maximization algorithms. In this paper, for the sake of Deepfake detection they have used ten famous and better generative Deepfake detection algorithms. As each of them vary in different aspects such as number of images generated, size, datasets.

For a given input image I, the authors propose extracting the description by calculating the local correlations. These calculations can be done via a convolution kernel as shown below:

$$I[x, y] = \sum_{s,t=-\alpha}^{\alpha} k_{s,t} * I[x+s, y+t] \qquad (1)$$

Generally, to find the Convolutional Traces, the Expectation Maximization algorithm [9] was used. As this algorithm helps in discrimination between the any two distributions that is the expected ones and the Deepfake images.

This includes the two steps:

- **Expectation Step:** Determines the probability whether it belongs to the model or not.
- **Maximization Step:** Probability estimations were done based on the instances.
- This process is done repeatedly.

Let us assume that P1 and P2 have different probability distributions. P1 be a Gaussian Distribution with zero mean and unknown variance and P2 is uniform. In the expectation step, the value of *I[x, y]* is calculated via Bayes Rule.

The result of EM is a feature vector that represents the structure of the Transpose Convolution Layers used to generate the image, and can be referred to as a Convolutional Trace (CT) because it encodes in some way whether the image is a Deepfake or not. Using Random Forest, the CT is used to distinguish real from Deep Fake pictures.

CT is applied on the image to extract a particular feature vector. It doesn't depend on training. Moreover, the CT isn't a Deep Learning Architecture. So, it cannot be applied for high level semantics.

Accuracy values are computed when 70% of the dataset is used for training and remaining is used for testing. Depending upon different kernel sizes, classifiers accuracy values are presented in the below table.

Table- 1: Accuracy values measured on different classifiers.

Classifier	*Kernel Size*	
	3x3	5x5
3-NN	89.80	77.38
SVM Linear	84.14	76.28
LDA	83.50	77.38
Random Forest	98.07	93.81

To conclude, they have presented a novel method for extracting convolutional traces based on the Expectation-maximization algorithm, which has strong discriminative power and is resistant to attacks.

3.2 Detecting Face Warping Artefacts

This paper [10] by Y Li, et al., looks at a deep learning and more specifically CNN based method to identify whether a particular image has been deepfaked or not. The method they propose is straightforward because it depends on the changes produced in the structure of the images as they are modified by deepfakes.

The primary idea behind their method is to detect the image artefacts [11], that is, irregularities in the image, which are produced by the process of deepfake. By detecting whether artefacts that are unique to the deepfake generation process are present in a particular image, one can tell whether the image is real or not.

For this particular purpose, the authors propose a neural network model which is composed of Convolutional Neural Networks (CNNs). They also choose to focus on those deepfake images whose faces are morphed via an affine transformation. The process of training the CNNs to detect goes thus.

1. Regions of interest (ROI) are extracted from positive and negative examples.
2. These ROI contain both the deepfaked face and the region surrounding it to expose the artefacts
3. The ROI is chosen to contain a bounding box for the face as follows -

[b_0-$\widehat{b_0}$, a_0-$\hat{a}_0$, b_1+$\widehat{b_1}$, a_1-$\hat{b}_1$], where a_0, a_0, b_1, b_1 denotes the minimum bounding box *B* which can cover all face landmarks excluding the outline of the skin. The variables b_0, a_0, b_1, a_1 are random value between [0, d/5] and [0, e/8], where d and e are the height and width of B respectively. The ROIs are resized to 224 × 224 to feed to the CNN models for training.

The authors trained 4 pre-defined CNN models, VGG16, ResNet50, ResNet101 and ResNet152. Each of these trained models is sent real deep fakes generated from various popular deepfake generators such as UADFV and DeepFakeTIMIT. Their results are shown below to facilitate better comparison.

Table- 2: AUC results of the CNN models trained to detect face warping artefacts

Methods	*UADFV*	*DeepFakeTIMIT*	
		Low Q.	*High Q.*
VGG16	84.5	94.6	57.4
ResNet50	97.4	99.9	93.2
ResNet101	95.4	97.6	86.9
ResNet152	93.8	99.4	91.2

We can infer from the results obtained that for any given deepfake image fed into these trained models, the ResNet50 gives the best performance in terms of accuracy. Further, one caveat given by the authors is that these detection methods primarily depend on the fact that deepfakes are currently being produced in a low image resolution. In the future, should they be produced at a resolution high enough to eliminate artefacts entirely, it would be tough to detect them via this method.

3.3 Deep-Fake Detection Using Visual Interpretability Methods

This paper [12] by Badhrinarayan Malolan, Ankit Parekh, Faruk Kazi is built on the pillars to safeguard human trust by detecting Deepfakes. Here, a new method is proposed which is based on Explainable Artificial Intelligence (XAI)[13, 14]. This is a deep learning where Convolutional Neural Networks were trained on FaceForensic dataset. Further, this is tested on Explainable Artificial Intelligence techniques such as LRP and LIME. Explainable Artificial Intelligence (XAI)helps interpret the result. Transparent intuitions can be obtained when one uses the approaches related to Machine Learning.

Traditional CNNs were used on the ImageNet database to conquer the feature from the images. Upon that Local Interpretable Model-Agnostic Explanations (LIME) are used [15]. LIME is most useful in interpretation of predictions from any classifier. It paves a path to extract the inferences. the equation for LIME is:

$$\xi(x) = argmin_{g\in G} l(f, g, \pi_x) + \Omega(g) \tag{2}$$

Layer-Wise Relevance Propagation (LRP)[16] is one of the testing methods in XAI. For each neuron a local redistribution rule is calculated. This produces pixel-wise distribution. Starting from the outer layer neurons to innermost layer neurons relevance scores are calculated and summed up to from equation as follows:

$$R_{i\leftarrow j} = \frac{z_i \times w_{ij} + \frac{\epsilon \times sign(z_j) + \delta \times b_j}{N}}{z_j + \epsilon \times sign(z_j)} \times R_j \tag{3}$$

$$R_j = \sum_j R_{i\leftarrow j} \tag{4}$$

Mechanism Involved:

1. Face extraction pipeline is created to comprise images of real and fake classes. These were divided into ratios 4:1:1 to train, test and validate.
2. For the sake of training Xception network comprising 134 layers is used. Binary cross-entropy function is used to meet the global minima. Test accuracy on different scales can be seen in table below:

Table -3: Sample of Results

Image Size	1.3x	2.0x
Test Accuracy	94.33%	90.17%

So, to know the true nature we can choose 2.0x image size. As the model performance can be calculated. Even robustness of the model can be inferred on Gaussian Blur Noise.

3. LIME was able to capture the critical areas around the face that led to its classification, as can be shown. LIME is capable of capturing the relevant parts of the image while ignoring unneeded background data. LIME responds effectively to affine transformations and Gaussian blur noise despite not having been trained on such pictures.

4. To adequately localize the manipulation regions to the nose and mouth region, LRP methods are applied to the input images. So, we can say that LRP's preserved the structural qualities.

In this paper, they have used XAI techniques as this helps to interpret the complex models. Models such as LIME and LRP were used to find remedies to the context which cannot be explored by AI. This helps us in better deepfake detection.

3.4 Deepfake Detection Based on Haar Wavelet Transform

This paper [17] by Mohammed Akram Younus et. al., proposes a scheme to detect whether an image is a deepfake or not by looking at the inconsistencies in the blur over the entire image. The authors aim to take advantage of a certain property of deepfake generation.

The basic parameters of deepfake generation such as the speed, limitation of resources, blur transformation and the affine transformation used on the changed images produce in its certain characteristics, which the authors have identified as an area of exploitation. The authors propose further to take advantage of the GAN transformations during deepfake generation and use those traces to identify the inconsistencies in blur and resolution in the Region of Interest.

The authors propose to detect such inconsistencies and aim to achieve this by means of an analysis done via applying a Haar Wavelet Transform [18]. The authors propose that an accurate method to detect inconsistencies would be to compare the blurred area in the ROI with the blur in the rest of the image. They further propose to perform this by use of a dedicated Haar Wavelet transform to simulate possible blurring and use that in their comparison.

The method proposed here uses the property of the Haar Wavelet Transform to discriminate between different edges and blurred image. and use this as the basis for detection. The further use the division of edges into Dirac Structures, Step Structure and Roof Structure to enhance the procedure.

1. The authors propose the following pipeline to detect deepfakes by using the Haar Transform as described above.
2. The faces and the ROI are extracted from the image.
3. Perform the Haar Wavelet Transform to decompose the image three times.
4. In each decomposition an edge map is constructed as follows with edge E being -
   $E\,map_i(K,l) \;=\; \sqrt{LH_i^2 + HL_i^2 + HH_i^2};\ (i = 1,2,3)$
5. By partitioning the edge maps local maxima in each window is found.
6. For a specific threshold value of the edge map, categorize it as an edge point, with the total number being in N_{edge}.
7. Similarly find all Dirac Structure edges and have them be N_{da}.
8. Let all the Roof Structure edges be N_{rg}.
9. Let N_{brg} be the total number of such roof structure edges which have lost sharpness as specified by a threshold.
10. Calculate ratio of all edges, per = N_{da}/N_{edge} and if per<a_{zero} then it signifies a blurred edge, where a_{zero} is a positive number close to zero
11. BlurExtent = $N_{brg/}N_{rg}$ to represent the image blur coefficient is calculated.
12. Compare the blur in ROI with the rest of the image if the blur is found in ROI.
13. From above comparison, we can determine whether the image is deepfaked or not.
    The authors of this paper have chosen to test their method against the UADFV dataset for deepfaked images and videos. Their results of the AUC are given in the table below for reference.

Table- 4: Comparison of AUC of different methods on UADFV

Methods	*UADFV AUC score*
Two stream NN	85.1
Meso-4	84.3
MesoInception4	82.1
HeadPose	89.0
Proposed Method	90.5

From this paper, we can look at a novel method of detecting deep fakes, something which doesn't involve neural networks directly and thus can be achievable with less overhead.

IV. CONCLUSION

In this paper, we present some recent advances in the field of deepfake detection. The methods presented here are by no means exhaustive but should offer a good starting point to anyone looking to solve the issue of deepfake detection or to get started in research in this domain. By presenting the techniques with a mathematical description as well as their accuracy on common deepfake datasets, a better idea of each has been provided. While each of these techniques alone might not be a solution to the deepfake problem, even an ensemble model developed from these is a future avenue to explore.

The problem of deepfakes and their prevalence presents an existential threat to trust and safety in a digital medium and it is our hope that the information presented in this review of the recent advancements in deepfake detection would help push research in this domain to newer venues. Deepfake generation and their subsequent detection is a cycle which is comparable to an arms race to develop more and more sophisticated techniques on each side. Therefore, it is of the utmost importance to make sure that newer methods are developed and catalogued with urgency. This paper is an effort to do the same. We hope that it has been of use in this regard.

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley,S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Advances in Neural Information Processing Systems 27, Montreal, Quebec, Canada, 2014, pp. 2672−2680.

[2] Omkar Metri, H.R Mamatha, Chapter 10 - Image generation using generative adversarial networks, Editor(s): Arun Solanki, Anand Nayyar, Mohd Naved, Generative Adversarial Networks for Image-to-Image Translation, Academic Press, 2021, Pages 235-262, ISBN 9780128235195

Deepfake and their Impact:

[3] A. A. Maksutov, V. O. Morozov, A. A. Lavrenov and A. S. Smirnov, "Methods of Deepfake Detection Based on Machine Learning," 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus), 2020, pp. 408-411, doi: 10.1109/EIConRus49466.2020.9039057.

[4] Buo, Shadrack Awah. "The Emerging Threats of Deepfake Attacks and Countermeasures." arXiv preprint arXiv:2012.07989 (2020).

[5] D. Yadav and S. Salmani, "Deepfake: A Survey on Facial Forgery Technique Using Generative Adversarial Network," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 852-857, doi: 10.1109/ICCS45141.2019.9065881.

[6] John Wojewidka, The deepfake threat to face biometrics, Biometric Technology Today, Volume 2020, Issue 2, 2020, Pages 5-7, ISSN 0969-4765

[7] L. Guarnera, O. Giudice and S. Battiato, "Fighting Deepfake by Exposing the Convolutional Traces on Images," in IEEE Access, vol. 8, pp. 165085-165098, 2020, doi: 10.1109/ACCESS.2020.3023037.

[8] L. O. Chua and T. Roska, "The CNN paradigm," in IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, vol. 40, no. 3, pp. 147-156, March 1993, doi: 10.1109/81.222795.

[9] T. K. Moon, "The expectation-maximization algorithm," in IEEE Signal Processing Magazine, vol. 13, no. 6, pp. 47-60, Nov. 1996, doi: 10.1109/79.543975

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

[11] Punchihewa, Amal, and Donald G. Bailey. "Artefacts in image and video systems; classification and mitigation." Proceedings of image and vision computing New Zealand. 2002.

[12] Malolan, A. Parekh and F. Kazi, "Explainable Deep-Fake Detection Using Visual Interpretability Methods," 2020 3rd International Conference on Information and Computer Technologies (ICICT), 2020, pp. 289-293, doi: 10.1109/ICICT50521.2020.00051.

[13] Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Information Fusion 58 (2020): 82-115.

[14] Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," in IEEE Access, vol. 6, pp. 52138-52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[15] Zafar, Muhammad Rehman, and Naimul Khan. "Deterministic Local Interpretable Model-Agnostic Explanations for Stable Explainability." Machine Learning and Knowledge Extraction 3.3 (2021): 525-541.

[16] Montavon, Grégoire, et al. "Layer-wise relevance propagation: an overview." Explainable AI: interpreting, explaining and visualizing deep learning (2019): 193-209.

[17] Younus, Mohammed Akram, and Taha Mohammed Hasan. "Effective and fast deepfake detection method based on haar wavelet transform." 2020 International Conference on Computer Science and Software Engineering (CSASE). IEEE, 2020.

[18] Stanković, Radomir S., and Bogdan J. Falkowski. "The Haar wavelet transform: its status and achievements." Computers & Electrical Engineering 29.1 (2003): 25-44.