Grenze International Journal of Engineering and Technology, Jan Issue

Deepfake Creation and Detection: A Review

Sunanda Dixit[1], Mahalakshmi J[2], Kirimanjeswara Anusree[3], Kumuda Siri T T[4] and Manisha M[5]
[1]Associate Professor, BMS Institute of Technology
[2-5]UG Student, BMS Institute of Technology

***Abstract*—Deep generative modeling has transpired into today's tech world probably more than ever before. Considering the complex datasets amalgamating a wide range of intricacies, deep generative models have still worked their way out in successfully generating results with maximum accuracy. With the advent of detection techniques, the advent of creation of fake contents have been proliferating in numerous ways. The escalation of such fake digital content has instilled skepticism in people, contributing to a zero-trust society where people no longer bother to distinguish truth from falsehood. Apart from the negative connotations associated with DeepFakes, their relevance has also been significant commercially as well as individually. Either way, this technology has triggered an increasing sense of unease in people.**

This concern proposes an urgent need for automated approaches to detect the fake content. Since these fake content are mostly produced from deep learning based approaches, they take up the name "DeepFakes". It can take up multiple forms such as images, videos or voice. Notwithstanding the fact that face editing algorithms have been able to yield convincingly realistic human faces, upon close scrutiny, have brought to limelight artifacts which are not visible to the naked eye. Both classical mathematical approaches as well as deep learning based approaches have been fairly successful in detecting these artifacts. In this work, we present a comparative analysis to distinguish which of the methods produce more accuracy in detecting Deepfake images. A classic approach involving frequency domain analysis(FDA) with a classifier is compared with a convolution neural network architecture with pros and cons respectively to infer which of the two methods produce clear-cut results.

I. INTRODUCTION

The face is the most idiosyncratic feature of a human being. Increasing sophistication of smartphones and social media have contributed to a massive pool of digital object content. Face synthesis is one domain which has been under constant experimentation leading to improved efficiency in every methodology discovered. With factors like time-consuming and domain expertise requirement, image alterations weren't prevalent. Thanks to the advances in deep learning and easy accessibility to large volumes of training data, these no longer pose a limitation. The aftermath of which made fabrications quite simple allowing amateurs to easily modify digital content to suit their needs. This efficiency is mostly not welcome since it brings along significant security risks.

Myriads of algorithms have made it possible to swap an individual's face with someone else's that appear strikingly genuine. Deepfake is one such domain of artificial intelligence technology which involves a certain face manipulation mechanism to alter contents of a digital space. Various techniques make the production of Deepfakes effortless. One is the use of autoencoders, a deep neural network, which uses a target video to collaborate a collection of video clips of the entity which must replace the entity in the target video. The video clips may be completely unrelated; the autoencoder makes use of the common features between both the entities

Grenze ID: 01.GIJET.9.1.537

Fig 1: Video frames from the Celeb-DF dataset where the first column corresponds to original video frames and the other frames correspond to synthesized faces

owing to various environmental conditions and angles to exactly merge in the two entities.. Another technique which has pushed through the stereotypes of the state-of-the-art methods and contributed to producing high realistic, high-resolution images is GANs[Generative Adversarial Networks]. It involves a collaboration between two competitive neural network models , a generator which learns to produce an output and a discriminator which learns to differentiate the true data from the output generated by the generator. The steps are looped through several times to achieve a good amount of accuracy[1].

Therefore, to eradicate, if not completely ,atleast to the largest extent possible the defamation, insecurities, scams and deception caused primarily by Deepfakes, researchers have been rigorously working on detecting DeepFakes.

There has been an upsurge in the detection mechanisms as well including both traditional machine learning classifiers. The main contribution of this work is to highlight various approaches delivering improved efficiency in terms of Deepfake detection and creation.

II. Literature Survey

A. Deepfakes

The term “Deepfake" derives its name based on the fact that it leverages deep learning technology to create fake digital content. Deepfakes have taken the spotlight these days to be more prevalent than ever before on the digital platform.Deepfake technology effectively manages to successfully transpose a face onto a target as if it were a mask. Presented below are a few papers addressing the increasing dominance of Deepfake technology.

Sl.No	AUTHOR NAME, YEAR	TITLE	METHODOLOGY	PROS	CONS
1.	Wang, Sheng et.al (2020) [1].	Cnn-generated images are surprisingly easy to spot... for now.	The images generated by CNN share few common systematic flaws, which increases the possibility of creating a generalized detector for deep fakes.	Even upon training on one GAN architecture, the model (ResNet50) is still able to generalize the unseen datasets and architectures.	The classifier model fails to achieve higher performance in case of CycleGAN, StarGAN, SAN. Training with more classes improves the performance but it is true only upto a point.

B. Deepfake detection using Frequency based methods:

One method of dealing with DeepFake digital content is utilizing frequency artifacts. Due to the unique appearance of frequency level artifacts, authentic and manipulated digital content generated from different GAN models are easily detected. Papers leveraging frequency-based detection to identify Deepfakes are presented below.

Sl.No	AUTHOR NAME, YEAR	TITLE	METHODOLOGY	PROS	CONS
1.	Durall, Ricard et al. (2019).[2]	Unmasking deepfakes with simple features.	Frequency domain analysis(FDA) is a strategy where the image in the spatial domain is converted to a frequency domain. Mathematical techniques namely Discrete Fourier Transform(DFT) and Azimuthal averaging are used to conceptualize a sturdy 1D representation of the FFT spectrum. Passed through a classifier such as SVM and Logistic regression to perform binary classification.	The analysis of the frequency domain followed by the binary classification utilizing simple traditional classifiers eliminates the need to involve vast quantities of data making it extremely feasible in situations such as data scarcity.	Compared to the input images posing different formats of uncertainties, the available frequency spectrum is smaller and hence identification of low resolution content is quite a task.
2.	Frank, Joel, et al.(2020).	Leveraging Frequency Analysis for Deep Fake Image Recognition	This method proposes a frequency-domain method of analysis of images [3] based on several artifacts which are common across different neural network architectures, datasets used and the resolution of images used for the process.	Performs an empirical comparison for detecting deepfake images. The classifiers yield a higher accuracy by considering few parameters.	The results are concerned only to one forensic classifier and are not adversarially robust.
3.	Jeong, Yonghyun, et al. (2022).	FrePGAN: Robust Deepfake Detection Using Frequency-level Perturbations	A framework is designed to generate a generalized deepfake detector. The framework [4] has a FrePGAN for generating the frequency-level perturbation maps and a classifier. Along with the perturbation generator, the deepfake detector is updated.	Reduces the effect of frequency level artifacts by generating perturbations. Puts forth a generalized detector which is not limited to training settings.	The performance is low compared to other models in BigGAN, CycleGAN and GauGAN. Increases with training catego
4.	Sun, Fang et al. (2021).	Deepfake Detection Method Based on Cross Domain Fusion	Makes use of FDA for extracting the edge geometric features. Haar transformation is used to extract the features in 3D [5].The features include the background, edges and corners of the image. Meso4, a neural network based architecture is used to get the spatial domain features that includes the specific features of the face.	The analysis is carried out by training using low quality data(i.e train dataset) and testing it with respect to high quality data(i.e test dataset).	Cannot recognize subtle changes in the features like lips, eyes, nose appropriately.
5.	Qian, Yuyang et al. (2020)	Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues	F3 -Net [6] makes use of the frequency-aware forgery clues is a method composed of two frequency-aware branches,the first branch focuses on mining the subtle forgery patterns and the other aims at extracting the small-scale discrepancy of frequency statistics that is observed between the real and forged face images.	The deepfaked artifacts whose visual quality is degraded cannot be captured in the RGB domain easily, but can be identified using the F^3-Net method.	A global feature based solution, F^3-Net does not work as expected as differentiating between fake and real faces has become more tenuous.

C. Deepfake creation and detection using CNN

Deep learning oriented generative models such as Generative Adversarial Networks(GANs) and variational autoencoders have been successful in generating photo-realistic, high-dimension digital content by collaborating with various other technologies. On similar lines, CNN architectures have also gone through multiple modifications to encounter various shortcomings leveraged by traditional CNN architectures and emerged to be successful in detecting Deepfakes. Papers addressing the usage of CNNs in tackling Deepfakes are presented below.

III. CONCLUSION

Deepfakes, due to its strong association with human faces, can be weaponized to function against the identity of an individual. The aftermath of which can be highly treacherous to the social, legal , financial and political fraternity. These intensified concerns have also triggered an upsurge in various detection methods. However,

Sl.No	AUTHOR NAME, YEAR	TITLE	METHODOLOGY	PROS	CONS
1	Hsu, C.C, et al. (2020)	Deep fake image detection based on pairwise learning	Incorporates a pairwise learning approach called common fake feature(CFFN)[7]. establishes a collaboration between Siamese network and DenseNet employing a semi-supervised learning strategy. A contrastive loss is computed to learn the discriminative features of the face termed as Common Fake Feature(CFF). The classifier adopts a cross-entropy function for learning.	Delivers better generalization. Can identify fake images even if they haven't been a part of the training phase. Outperforms the other state-of-the-art methods in terms of precision and recall rate.	Retrains the image detector since the CFFs are learned from pairwise learning samples. May fail when there is a difference between the fake features produced by the new generator and those produced in the training phase.
2	Zhou, P et al. (2017).	Two-stream neural networks for tampered face detection	Hybrid approach based on two-stream fusion [8]. GoogleNet, is embodied in the model to train on tampered and authentic images. The other stream is a patch triplet stream which comprises a steganalysis feature extractor with a triplet loss, facilitating analysis of features. To obtain better insights on what face classification streams can provide visually, Class Activation Maps(CAMs) have been leveraged.	Better performance compared to individual stream by a margin. Different noise residual patterns are also identified utilizing patch triplet network in the tampered region.	Upsample and resizing the face input to 299*299 causes chances of losing visual information. Patch size requirement of 128*128 renders the approach less robust since a major portion of the image may be authentic.
3.	Amerini, I at al. (2019)	Deepfake Video Detection through Optical Flow based CNN	Optical flows exploit disparities in the motion across the frames of artificially and naturally generated videos. A forward flow for each frame, at a certain time,is extracted using a CNN model named PWC-Net[9] . This calculation is provided as an input to the semi-trainable CNN called Flow-CNN.	The optical flow fields utilizes the feasible inter-frame dissimilarities in contrast to the state-of-art techniques that resorts at single video frames.	Models generated by Optical flows are computationally complex and are highly noise sensitive.
4.	Nguyen, T.T, at al. (2019).	Deep learning for deepfakes creation and detection: A survey	Gaussian blur and Gaussian noise remove high frequency clues thereby enhancing generalization. Bag of words strategy used to extract a set of compact features and later fed into classifiers such as multi-layer perceptrons [10]. The deep network recurrent models identify temporal patterns across video frames. Block chain technologies and smart contracts have been leveraged with the primary assumption that videos exist only when their source is traceable.	Storage of video metadata in a decentralized fashion. Tested against security challenges like replay, man-in-the-middle and DOS attacks. Can be applied to all digital contents	Experimentations on same-forgery and in-dataset values caused incapability in generalization. Adversarial perturbation attacks have been successful in fooling DNN based detectors making deepfake detection infeasible.
5.	Chang, X et al. (2020).	DeepFake face image detection based on improved VGG convolutional neural network.	An improvised VGG network named NA-VGG[11] detects deepfakes. The input to the network is the image noise which is obtained by passing it through a SRM filter layer followed by image augmentation which can flip,translate,rotate or zoom the image, thus enhancing the ability of existing data to generate new images suitable for training and testing.	Image noise feature and augmentation improves the efficiency of detection without restricting to only color mismatch and visible boundary of images.Reduces the problem of overfitting.	The weights of the SRM filter kernel in noise flow is 5*5*3. Employs only horizontal or vertical image augmentation. Every image has to be resized to 128*128.
6.	Shad, H.S, et al.(2021)	Comparative Analysis of Deepfake Image Detection Method Using Convolutional	Eight different deep learning architectures including DenseNet, VGG16, VGG19, VGGFace, CustomCNN were evaluated against five different evaluation metrics, including accuracy, precision, F1-score, recall, and	Amidst all the eight different CNN models, the performance of VGGFace was best with an accuracy of	Even though VGGFace has 99% accuracy, with respect to recall, the customCNN model performance was better.

		Neural Network	the area under the ROC curve [12].	99%.	
7.	Koopman, M, et al.(2018)	Detection of deepfake video manipulation.	Identifies the visual artifacts using Photo response non uniformity (PRNU) analysis [13]. The frames are then batched in groups of 8 having equal size and a PRNU pattern is obtained.Each of these are compared with 7 other patterns and a mean normalized cross correlation score is determined.	The authentic videos and the deep fakes are easily distinguished using the calculated normalized cross correlation score.	The dataset considered is very small to draw conclusions about the performance and the evaluation of the stated method.
8.	Do, N.T et al. (2018).	Forensics Face Detection From GANs Using Convolutional Neural Network, 2018	A deep convolutional neural network model which works based on Deep Feature extraction in the VGGFace network architecture[14] is proposed.	DNNs have the advantage of automated learning and feature extraction.	Only focuses on the images generated using GAN and not other types of inputs.
9.	Passos, L.A, et al.(2022).	A Review of Deep Learning-based Approaches for Deepfake Content Detection	A review on several deep-learning based approaches used in detecting the deepfake content[15], which covers the advantages, disadvantages,shortcomings and steps to solve them.	More dynamic methods should be used for detecting the deepfakes generated using GANs.	Effectiveness of the model with and without fine-tuning procedure must be analyzed.
10	Cozzolino, D, et al. (2021).	ID-Reveal: Identity-aware DeepFake Video Detection	Biometric features are used to detect manipulated video content. The temporal ID Network and the 3DMM Generative Network are employed [16]. Compact representation of each frame of the input video is extracted using the 3DMM and these features are given as an input to the Temporal ID network, the embedded vector is computed. Using Temporal ID as a discriminator, the 3DMM is trained to ensure that Temporal ID Network is based on behavioral characteristics.	The ID-Reveal technique is more generalized and is more vigorous to low quality videos compared to the state-of-art methods.	They rely on video only features, sometimes accompanied by categorical information. Thus it neglects the audio information. Moreover, in order to work properly, they need several hours of videos.

considering the high stakes that they carry along, there are diverse critical problems that need to be addressed by current Deepfake detection methods. Envisioning this aspect, the future scope of deepfake detection must majorly focus on certain directions. Face synthesis owing to increasing sophistication of generation algorithms must be addressed. This must include pivotal aspects such as head puppetry and lip syncing. Impersonations have not only been restricted to imagery, high-quality audios and videos have also escalated quickly. Since audio signals are 1D in nature, image and video detection methods may not be applicable to audios. However robust voice spoofing detection frameworks have addressed this issue [17]. Forensic methods need to be involved to demarcate contents produced with malicious intent and contents produced with satirical intent [18]. Deepfake creation methods still stand foible. Minuscule details such as skin and facial hair haven't been addressed due to the loss of information in the encoding step. However studies suggest incorporating GAN models deliver promising performance [19][20]. Future research in these directions highlights a promising path for Deepfake creation and detection to coexist in the society.

REFERENCES

[1] Wang, Sheng, Y, W Oliver, Z Richard, O Andrew, and A. E Alexei. "Cnn-generated images are surprisingly easy to spot... for now." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020. pp. 8695-8704.

[2] Durall, Ricard, K Margret, P Franz-Josef, and K. Janis. "Unmasking deepfakes with simple features." arXiv preprint arXiv:1911.00686., 2019.

[3] Frank, Joel, Thorsten Eisenhofer, Lea, Schönherr, Fischer Asja, Kolossa Dorothea, and Thorsten Holz. "Leveraging frequency analysis for deep fake image recognition." In International Conference on Machine Learning. PMLR, 2020. pp. 3247-3258.
[4] Jeong, Yonghyun, Kim Doyeon, Ro Youngmin, and Choi Jongwon. "FrePGAN: Robust Deepfake Detection Using Frequency-level Perturbations." arXiv preprint arXiv:2202.03347, 2022.
[5] Sun, Fang, Z Niuniu, X Pan, and S Zengren. "Deepfake Detection Method Based on Cross-Domain Fusion." Security and Communication Networks, 2021.
[6] Qian, Yuyang, Yin Guojun, Sheng Lu, Chen Zixuan, and Jing Shao. "Thinking in frequency: Face forgery detection by mining frequency-aware clues." In European Conference on Computer Vision. Springer, Cham, 2020. pp. 86-103.
[7] Hsu, C.C, Y.X Zhuang, and C.Y Lee. "Deep fake image detection based on pairwise learning." Applied Sciences, 10(1), 2020: p.370.
[8] Zhou, P, X Han, V.I Morariu, and L.S Davis. "Two-stream neural networks for tampered face detection." In 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW). IEEE, 2017. pp. 1831-1839.
[9] Amerini, I, L Galteri, R Caldelli, and A Del Bimbo. "Deepfake video detection through optical flow based cnn." In Proceedings of the IEEE/CVF international conference on computer vision workshops. 2019..
[10] Nguyen, T.T, Q.V.H Nguyen, C.M Nguyen, D Nguyen, D.T Nguyen, and S., Nahavandi. " Deep learning for deepfakes creation and detection: A survey." arXiv preprint arXiv:1909.11573, 2019.
[11] Chang, X, J Wu, T Yang, and G Feng. "DeepFake face image detection based on improved VGG convolutional neural network." In 2020 39th chinese control conference (CCC) . 2020. pp. 7252-7256.
[12] Shad, H.S, et al. "Comparative Analysis of Deepfake Image Detection Method Using Convolutional Neural Network." Computational Intelligence and Neuroscience, 2021
[13] Koopman, M, A.M Rodriguez, and Z Geradts. "Detection of deepfake video manipulation." In The 20th Irish machine vision and image processing conference (IMVIP). 2018. pp. 133-136.
[14] Do, N.T, I.S Na, and S.H Kim. "Forensics face detection from GANs using convolutional neural network." ISITC. 2018. pp.376-379.
[15] Passos, L.A, D Jodas, K.A da Costa, L.A.S Júnior, D Colombo, and J.P Papa. "A Review of Deep Learning-based Approaches for Deepfake Content Detection." arXiv preprint arXiv:2202.06095., 2022.
[16] Cozzolino, D, A Rössler, J Thies, M Nießner, and L Verdoliva. "Id-reveal: Identity-aware deepfake video detection." In Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021. pp. 15108-15117.
[17] Chen, Tianxiang, and et al. "Generalization of audio deepfake detection." Proc. Odyssey 2020 The Speaker and Language Recognition Workshop, 2020.
[18] Verdoliva, and Luisa. "Media forensics and deepfakes: an overview." IEEE Journal of Selected Topics in Signal Processing 14.5, 2020: 910-932
[19] Karras, Tero, and et al. "Progressive growing of gans for improved quality, stability, and variation." arXiv preprint arXiv:1710.10196, 2017.
[20] Karras, Tero, Laine Samuli, and Aila Timo. "A style-based generator architecture for generative adversarial networks." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.