Emerging Techgaurd on Deepfake detection

Gurleen Kaur
Department of CSE,
Chandigarh University,
Punjab, India
khushigurleen23@gmail.com
Vigya Vigy
Department of CSE,
Chandigarh University,
Punjab, India
Vigyavigy555@gmail.com

Anshika Mittal
Department of CSE,
Chandigarh University,
Punjab, India
mittalanshika28@gmail.com

Adil Husain Rather
Department of CSE,
Chandigarh University,
Punjab, India
adilhusain5057@gmail.com

***Abstract*—Deepfake technology has been undoubtedly growing at a rapid pace since 2017. Specifically, GAN model and other techniques related to it were majorly popularized at that time. The recent trends of emerging artificial intelligence and deep learning models make it easier for the forgers to create fake images, audio, and videos. While manipulating visual and auditory media is a long-standing tactic in deception, the emergence of deepfakes has signaled a sea change in the production of false content. With the development of artificial intelligence and machine learning, it is now possible to generate fake content using automated processes that are much more difficult for the average person to distinguish from the original. As Deepfake technology advances, so do the techniques to detect them like GAN's (Generative Adversarial Networks). Other deep learning approaches are being developed to counteract GAN-generated deep fakes. This paper particularly looks at the research done in this area by various other researchers and engineers. It also focusses on research that can influence it by looking at papers regarding human pose transfer, human motion transfer, and human motion generation. Finally, the paper highlights the promising directions and future research opportunities in the field of deepfake detection. As the arms race between deepfake generators and detectors continues, collaborative efforts from academia, industry, and policymakers are crucial to developing robust defenses against the misuse of deepfake technology.**

Keywords- Deepfake, AI, ML, GAN's, Deep learning, human motion transfer.

I. INTRODUCTION

In the era of advanced artificial intelligence and deep learning, the term "deepfake" has emerged as both a technological marvel and a growing concern. In the past two decades, the trend of creating fake media using deepfake has increased rapidly. This makes use of deep learning models, particularly GAN's. "The 21st centuries' answer to photoshopping, deepfakes use a form of artificial intelligence called deep learning to make images of fake events, hence the name deepfake (deep learning + fake)."

Some of the wrong uses of deepfakes are pornography (with the image of popular actresses), political speeches where the audio of a particular person can be used over the face/video of another person using the lip-syncing technology or similar thing can be done for religious speeches. Deepfakes were first introduced into the world when some adult video clips were tampered with the face of a popular personality. After this incident, the reddit community had an explosion of fake content. Another instance is when former president Barack Obama's speech was tampered with. The primary concern surrounding deepfakes lies in their potential for fraud. This deception can manifest in various ways, from spreading false information and hoaxes to perpetrating identity theft and financial fraud. As a result, deepfakes have garnered attention not only for their creative potential but also for their capacity to undermine trust and manipulate public discourse.

In an age where the boundaries between reality and fiction are increasingly blurred, understanding deepfakes is essential for society to navigate the challenges and opportunities they present.

II.RELATED WORK

N Choi et al (2023) with no requirement for artificial intelligence (AI), this research provides a ground-breaking blockchain-based technique for identifying online movie fraud. It responds to the expanding demand for impartial and reliable detection methods in the face of developing video content by combining user contributions and cutting-edge algorithms to enable transparent and reliable detection of fake movies. In comparison to public blockchains built on PoW, Hyperledger Fabric offers efficiency benefits for service provider agreements. Significant study was done on smart contract design and testing by Sánchez-Gómez et al. [16] and Górski [17], who emphasized the value of methodical consideration and verification techniques throughout the software development life cycle. [1]

Rajneesh Rani et al (2022) discussed that deepfake is a method for producing fake faces that may be used in place of real ones in photos or movies. It is based on machine learning and artificial intelligence. General Adversarial Networks (GANs) are the key element that makes deep fakes more realistic than ever; with the aid of these network systems incredibly convincing deep fakes can be formed. Although many different strategies have been developed in this field, the majority of them fall under the headings of facial artifacts, neural networks, and 3D head position. [2]

B Xu et al (2021) discussed that Deep learning is currently used to identify Deepfake films, which build deep neural networks to identify the frame sequence following framing. Recurrent Neural Network (RNN)-based detection is suggested in [9] based on the observation that some inconsistent choices, such as illuminants, exist across scenes containing fake frames. Convolutional Neural Network (CNN) is used to extract face information after first dividing the movie into frames. Long Short-Term Memory (LSTM) network is a collection of features from the videos' eye blinking detection and sent into an LSTM network to identify the subjects' eyes [18]. The results of the experiments demonstrate that the suggested method may successfully identify Deepfake films.[3]

Tao Zhang et al (2022) demonstrated that recent improvements in deepfake production make it more realistic, which makes it challenging to detect. The term "facial reenactment" refer to realistic-looking but false images, sounds, and movies produced by AI techniques. Deep fake has posed a serious danger to national security, democracy, society, and privacy, necessitating the use of existing datasets and detection techniques. Deepfake identification presents some difficulties, such as a dearth of good standards and datasets. [4]

Lakshmanan Nataraj et al (2019) discussed that deepfakes, image-to-image translations, and other automated techniques using GAN have grown incredibly popular. In order to detect GAN generated false images, a mix of co-occurrence matrices and deep learning has been used in this paper. Co-occurrence matrices for each of the three-color channels are gathered in the pixel domain, and a deep convolutional neural network (CNN) architecture is employed to train a model. The use of GANs in steganalysis to detect the existence of concealed data in an image has received considerable research. The co-occurrence technique has historically relied on a matrix and a machine-leaning classification algorithm like SVM. [5]

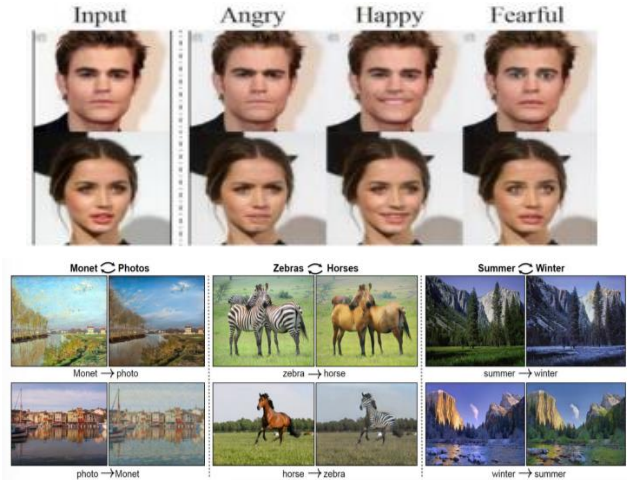

Figure 1: Examples of images generated using GAN.[5]

Joseph Bamidele Awotunde et al (2022) investigates privacy issues and the propagation of false information caused by the pervasive usage of deepfake technology, which creates fake videos. This research introduces a deepfake fake model that runs effectively by utilizing a five-layered Convolution Neural Network (CNN) technique. A novel approach to deep fake detection using a specialized CNN architecture strengthened by ReLU activation functions. In order to address new issues, it also acknowledges the dynamic nature of deepfake technology and emphasizes the urgent need for continual improvements in deepfake detection. [6]

Trisha Mittal et al (2020) discussed that main method used in this study to spot deep fakes is to take advantage of how the audio and visual components of a video source interact with one another. Prior studies, including those in psychology and multimodal machine learning [18], have repeatedly shown a high correlation between different modalities related to the same topic. There is a positive link between audio and visual modalities, a relationship that has been useful in the field of multimodal perceived emotion identification. This results from the various ways that emotions are expressed, which may cause our method to find discrepancies in the modalities of real videos and, as a result, incorrectly classify them by labeling them as false. [7]

Yuezun Li et al (2018) discussed that deepFake technology mainly relies on training deep neural networks with facial photos. In this research, we provide a revolutionary deep learning-based system that can distinguish between DeepFake films and real ones with great accuracy. Our method takes advantage of a distinctive feature of DeepFake videos: the DeepFake algorithm can only produce facial images of a limited size. As a result, these synthesized faces must go through an affine warping procedure to match the source face's arrangement. Because the warped face area and the surrounding context have different resolutions, this warping procedure generates obvious artifacts. These artifacts work as helpful markers for spotting DeepFake films. By employing a customized Convolutional Neural Network (CNN) model to compare the generated facial regions with their surrounding areas, our method can detect these artifacts.[8]

Y Li et al (2019) discussed that we provide Celeb-DF3, a brand-new, sizable dataset created particularly to test and analyse DeepFake detection algorithms. More than 2 million frames from 5,639 DeepFake videos make up this dataset. These videos are contrasted with true source videos that came from publicly accessible sources. 59 celebrity YouTube videos reflecting a range of ages, genders, and ethnicities. When compared to previous datasets, the visual quality of the DeepFake movies in Celeb-DF3 is noticeably better. The findings highlight the substantial obstacle Celeb-DF3 presents to most current detection techniques. [9]

T Liang et al (2020) discussed a novel method for efficiently capturing a variety of face emotions, head positions, and complex situations in movies using discrete, large interval sampling. This approach seeks to reach thorough conclusions about a video's veracity, especially in situations when real frames are intermingled with edited content. We introduce the SDHF, a hierarchical framework that uses 2D convolutional neural networks to extract frame-level data. Then, clip-level and video-level features are extracted using a 1D convolutional aggregator. SDHF Framework combines elements at three separate levels—frame, clip, and video—allowing for a thorough evaluation of video authenticity. The Best Sampling Method enables the thorough sampling of video situations. Assessment Using Multiple Datasets evaluates the SDHF approach's effectiveness using the DFDC, Celeb-DF, Face-Forensics++, and UADFV datasets.[10]

III. METHODOLOGIES

1. *Entire face synthesis*

Its goal is to use a neural network $\varphi(\cdot)$ to construct an imaginary face image xf from a random vector v.

That is $xf = \varphi(v)$.

Neural networks that are useful for full-face synthesis tasks are GANs and VAEs. PGGAN, StyleGan, and other methods are a few of the widely used complete face synthesis techniques. It has been demonstrated that these methods yield excellent deepfake photos. When compared to GANs, the images produced by VAEs are somewhat blurry. These blurry visuals are mostly caused by the training principle, which gives VAEs a high chance. This method introduces GAN related studies primarily because the images generated by VAEs are not realistic enough.

The mapping from a random distribution—typically Gaussian noise—to a distribution of human faces is learned by GANs. Their goal is to produce lifelike visuals that are hard to tell apart from actual human faces.

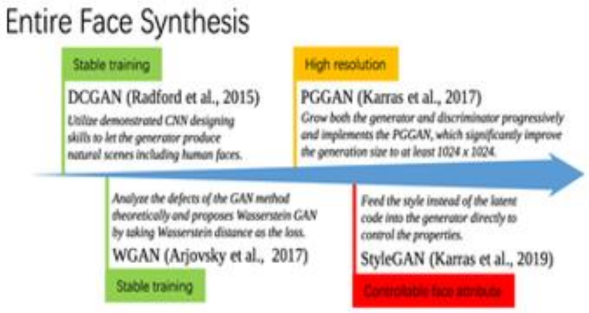
Entire Face Synthesis
Stable training
DCGAN (Radford et al., 2015)
Utilize demonstrated CNN designing skills to let the generator produce natural scenes including human faces.
High resolution
PGGAN (Karras et al., 2017)
Grow both the generator and discriminator progressively and implements the PGGAN, which significantly improve the generation size to at least 1024 x 1024.
Analyze the defects of the GAN method theoretically and proposes Wasserstein GAN by taking Wasserstein distance as the loss.
WGAN (Arjovsky et al., 2017)
Stable training
Feed the style instead of the latent code into the generator directly to control the properties.
StyleGAN (Karras et al., 2019)
Controllable face attribute

Fig2: Entire face synthesis[14]

2. *Attribute manipulation definition*

Its goal is to use neural networks $\varphi(\cdot, \cdot)$ to create a new fake image xf by altering the facial features P of an actual face image xr.

That is $xf = \varphi(xr, P)$.

The GANs for attribute manipulation can accomplish more precise attribute management using an improved attribute disentangle method. Current cutting-edge techniques, such as HifaFace (Gao et al., 2021d), may precisely alter facial features without sacrificing the rich details of non-editing regions.

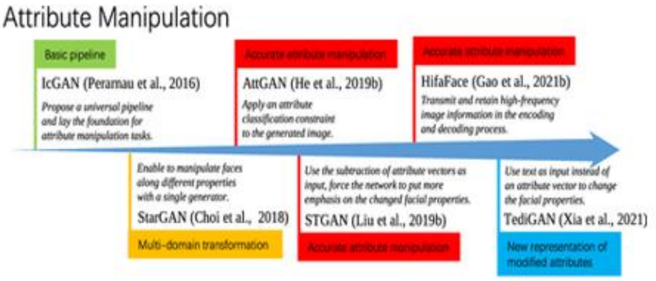
Attribute Manipulation
Basic pipeline
IcGAN (Peramau et al., 2016)
Propose a universal pipeline and lay the foundation for attribute manipulation tasks.
Accurate attribute manipulation
AttGAN (He et al., 2019b)
Apply an attribute classification constraint to the generated image.
Accurate attribute manipulation
HifaFace (Gao et al., 2021b)
Transmit and retain high-frequency image information in the encoding and decoding process.
Enable to manipulate faces along different properties with a single generator.
StarGAN (Choi et al., 2018)
Multi-domain transformation
Use the subtraction of attribute vectors as input, force the network to put more emphasis on the changed facial properties.
STGAN (Liu et al., 2019b)
Accurate attribute manipulation
Use text as input instead of an attribute vector to change the facial properties.
TediGAN (Xia et al., 2021)
New representation of modified attributes

Fig3: Attribute Manipulation[14]

StarGAN (Choi et al. 2018) and selective transfer GAN (STGAN) (He et al. 2019b) are two classical examples. Perarnau et al. (2016) introduced Invertible Conditional GAN (IcGAN), the first attempt at manipulating facial attributes using GANs.

3. *Identity swap definition*

Identity swap uses a neural network $\varphi(\cdot,\cdot)$ to create a new fake image xf by replacing the identity of the source image xs with the identity ti of the target image xt.

That is $xf = \varphi(xs,ti)$.

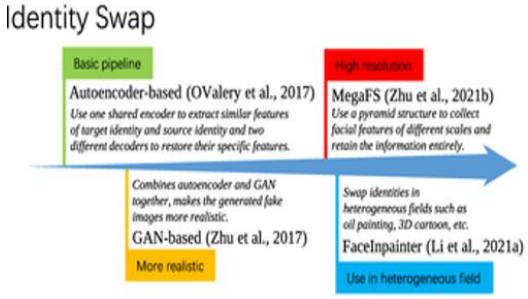
Identity Swap
Basic pipeline
Autoencoder-based (OValery et al., 2017)
Use one shared encoder to extract similar features of target identity and source identity and two different decoders to restore their specific features.
High resolution
MegaFS (Zhu et al., 2021b)
Use a pyramid structure to collect facial features of different scales and retain the information entirely.
Combines autoencoder and GAN together, makes the generated fake images more realistic.
GAN-based (Zhu et al., 2017)
More realistic
Swap identities in heterogeneous fields such as oil painting, 3D cartoon, etc.
FaceInpainter (Li et al., 2021a)
Use in heterogeneous field

Fig4: Identity swap[14]

Improvements in image quality have been made in the field of identity swapping in creating fake videos, mostly using autoencoder-based and GAN-based architectures. Autoencoder-based techniques enable the extraction of both general and specialized features by using a shared encoder and two independent decoders trained with source and target identities. The problem of paired examples has been solved by GAN-based approaches, particularly CycleGAN, by creating mappings between source and target domains, maintaining cycle consistency. FaceShifter pioneered a two-stage, high-fidelity face-swapping process that placed an emphasis on using the target image's adaptive information. To provide effective appearance mapping, appearance optimal transport (AOT) reduced the Wasserstein distance of latent characteristics. MegaFS accomplished one-shot ultra-high-resolution face swapping whereas InfoSwap concentrated on disentangling identity representation. Realistic switched faces with fused borders are produced by controllable face inpainting over a variety of domains using well-designed and refined networks.

IV. PROPOSED SOLUTION RESULT AND ANALYSIS

As discussed earlier, GAN’s are widely used technologies in making deepfakes and they are one of the most accurate models available till now. Images and videos created by GAN’s are closely related to the original media. Since there is no 100% accurate model to detect deepfakes yet, a combination of two models is generally preferred. In this paper, we will be working on DCGAN (an architecture of GAN) and CNN.

Preparing the model:

Step 1: The initial stages include the pre-processing of the data. This step involves importing a dataset of images and videos into a framework. Next, the data is split into training and testing sets, which are used to, respectively, train and evaluate the model.

Furthermore, it is preferable to resize the images to a standard size in order to increase the model's accuracy.

Step 2: Next step is to create two models: the generator to generate fake images and the discriminator to evaluate real and fake images. Now, the DCGAN model is trained by feeding it with real and convincing fake images.

DCGAN loss function: It measures how well the discriminator can discern between authentic and fraudulent images and how well the generator can produce realistic images. Reduce this loss as much as possible to raise the caliber of photographs that are generated.

$DCGAN_{loss}= G_{loss}+D_{loss},$ *(G is generator and D is discriminator)*

Where, $G_{loss} = -log(D(G(z)))$

$D_{loss} = -log(D(x))-log(1-D(G(z)))$

(x is real image and z is noise vector)

Step 3: Now that the generator model can produce fake images and it is saved now it’s time to define CNN model to extract features from images. The CNN model is trained using both the fake images created by DCGAN generator and real images from the loaded dataset. It trains the model to distinguish between them.

After this, CNN model extracts features from images used for image classification. CNN loss function measures how well CNN model performs in distinguishing.

$CNN_{loss} = -Y*log(p)-1(1-Y)*log(1-p)$

X is input image, Y is generated image and p is predicted probability

Step 4: In the final step of preparation of the model, we now combine the features extracted by CNN and DCGAN model. The goal of this information fusion is to make use of both the CNN's feature learning skills and the DCGAN's image generating capabilities. Now, the combined features are served as an input to a new layer of classifier. This layer makes classification prediction.

The model is trained with these combined features to recognize patterns.

CNN combined loss function quantifies the performance of the combined model in its classification task.

$CNN_{loss} = -Y*log(p)-1(1-Y)*log(1-p)$

X is input image, Y is generated image and p is predicted probability

Model name	Accuracy	Precision	Recall	F1 score
CNN	0.95	0.96	0.94	0.95
GAN	0.92	0.93	0.90	0.91

Table: accuracy measure.

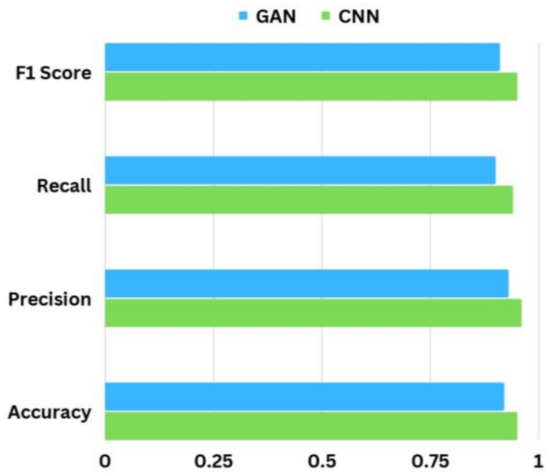

Fig5: Vsualization of accuracy measure.

V. FUTURE WORK

Although, Deepfakes have been into the market since almost two decades, but still there is a lot of research work to be done on this topic as we do not have any accurate foolproof method to detect manipulated media using deepfakes. Some CNN and machine learning algorithms have proved to be very helpful in achieving our target but none of them is 100% reliable. So, here are some of the works that can be done:

1. Development of more advanced and accurate deepfake detection method. This might involve the use of more sophisticated machine learning algorithms, combining multiple modalities (e.g., audio and video), and exploring novel features or representations.
2. Researchers can focus on making these models more transparent and understandable, helping users trust the decisions made by these systems.
3. Deepfake technology is extending beyond images and videos to generate text-based content. Future work can delve into detecting fake text generated by GANs to combat disinformation and fake news.
4. Developing real-time deepfake detection systems is essential, especially in applications where immediate action is required, such as in live-streamed content or video conferencing.
5. Deepfake techniques are continually evolving. Researchers should work on developing detection methods that can generalize to detect deepfakes created using new technologies and approaches.
6. Creating diverse and comprehensive datasets for deepfake detection is crucial. Future research can focus on curating more challenging and realistic datasets for benchmarking detection methods.
7. Developing user-friendly tools that enable individuals to check the authenticity of media content easily can empower users to protect themselves from deepfake-related threats.

Future work in the field of deepfakes should not only focus on detection but also encompass prevention, mitigation, and the ethical considerations surrounding this technology. It requires a multidisciplinary approach that combines expertise in machine learning, computer vision, cybersecurity, law, ethics, and public awareness.

VI. CONCLUSION

The conclusion of this research paper summarizes what is deepfake - "The 21st centuries' answer to photoshopping, deepfakes use a form of artificial intelligence called deep learning to make images of fake events, hence the name deepfake (deep learning + fake)." the various ways in which deepfakes are used to create and spread forged media on internet and how it is exploited in different ways like pornography, religious and political speeches. We have also discussed different research and study that has been conducted on deepfakes and methods to identify them. In this paper, we have also created the model using CNN and GAN to identify deepfakes and discussed about its results and analysis of features and their limitations.

VII. REFERENCES

[1] Choi, N. and Kim, H., 2023. DDS: Deepfake Detection System through Collective Intelligence and Deep-Learning Model in Blockchain Environment. *Applied Sciences*, *13*(4), p.2122.

[2] Rani, R., Kumar, T., Sah, M.P. (2022). A Review on Deepfake Media Detection. In: Sharma, H., Shrivastava, V., Kumari Bharti, K., Wang, L. (eds) Communication and Intelligent Systems . Lecture Notes in Networks and Systems, vol 461. Springer, Singapore.

[3] Xu, B., Liu, J., Liang, J., Lu, W. and Zhang, Y., 2021. DeepFake Videos Detection Based on Texture Features. *Computers, Materials & Continua*, *68*(1).

[4] Zhang, T. Deepfake generation and detection, a survey. *Multimed Tools Appl* **81**, 6259–6276 (2022). https://doi.org/10.1007/s11042-021-11733-y

[5] Nataraj L, et al (2019) Detecting GAN generated fake images using co-occurrence matrices. arXiv:1903.06836

[6] Awotunde, J.B., Jimoh, R.G., Imoize, A.L., Abdulrazaq, A.T., Li, C.T. and Lee, C.C., 2022. An Enhanced Deep Learning-Based DeepFake Video Detection and Classification System. *Electronics*, *12*(1), p.87.

[7] Mittal, T., Bhattacharya, U., Chandra, R., Bera, A. and Manocha, D., 2020, October. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 2823-2832).

[8] Li, Y. and Lyu, S., 2018. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*.

[9] Li, Y., Yang, X., Sun, P., Qi, H. and Lyu, S.C.D., 1909. A Large-scale Challenging Dataset for DeepFake Forensics (2019). *URL http://arxiv. org/abs/1909.12962*, *35*, p.36.

[10] Liang, T., Chen, P., Zhou, G., Gao, H., Liu, J., Li, Z. and Dai, J., 2020, November. SDHF: spotting deepfakes with hierarchical features. In *2020 IEEE 32nd international conference on tools with artificial intelligence (ICTAI)* (pp. 675-680). IEEE.

[11] Liy, C.M. and InIctuOculi, L.Y.U.S., 2018, December. Exposingaicreatedfakevideosbydetectingeyeblinking. In *Proceedings of the 2018 IEEE International workshop on information forensics and security (WIFS), Hong Kong, China* (pp. 11-13).

[12] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A. and Schulman, J., 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, pp.27730-27744.

[13] Baltrušaitis, T., Ahuja, C. and Morency, L.P., 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, *41*(2), pp.423-443.

[14] Baccarella, C. V., Wagner, T. F., Kietzmann, J. H., & McCarthy, I. P. (2018). Social media? It's serious! Understanding the dark side of social media. European Management Journal, 36(4), 431-438.