ISSN 2278-3091
Volume 9, No.3, May - June 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse58932020.pdf
https://doi.org/10.30534/ijatcse/2020/58932020

Deepfake Forensics, an AI-synthesized Detection with Deep Convolutional Generative Adversarial Networks

Dafeng Gong[1, 2*], **Ong Sing Goh**[3], **Yogan Jaya Kumar**[4], **Zi Ye**[5], **Wanle Chi**[6]
[1]Department of Information Technology, Wenzhou Polytechnic, Wenzhou 325035, China, 289133894@qq.com
[2]Centre for Advanced Computing Technology, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka 76100, Malaysia, 289133894@qq.com
[3]Centre for Advanced Computing Technology, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka 76100, Malaysia, osgoh88@gmail.com
[4]Centre for Advanced Computing Technology, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka 76100, Malaysia, yogan@utem.edu.my
[5]Centre for Advanced Computing Technology, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka 76100, Malaysia, yezi1022@gmail.com
[6]Centre for Advanced Computing Technology, Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Melaka 76100, Malaysia, 358455713@gmail.com
Corresponding author: Dafeng Gong (289133894@qq.com).

ABSTRACT

Recently, artificial intelligence, deep learning and Generative Adversarial Networks (GANs) adaptabilities for deepfake detection and forensics have become an emerging field of research interest. GANs have been widely studied since it was first proposed, and many applications have been produced to generate contents such as videos and images. The application of these new technologies in many fields makes it more and more difficult to distinguish between true and fake content. This study analyzes more than hundred published papers related to the application of GANs technology in various fields to generate digital multimedia data and expounds the technologies that can be used to identify deepfakes, the benefits and threats of deepfake technology, and how to crack down deepfakes. The findings indicate that although deepfakes pose a major threat to our society, politics and commerce, a variety of means are listed to limit the production of unethical and illegal deepfakes. Finally, the study also puts forward its limitations and possible future research directions and recommendations.

Key words: Artificial Intelligence, Deep Learning, Deepfake, Forensics, GANs

1. INTRODUCTION

For the last several years, with the development of Artificial Intelligence (AI) technologies through deep learning, automatic content generation has made remarkable progress, and the quality and form of content have been improved. The deepfake technology can realize the tampering, forgery and automatic generation of images, sounds and videos, and produce highly realistic and difficult to distinguish effects. Fake information has become a serious threat to the public news, democracy and human society[1]. The most popular product of deepfake is AI face changing technology, such as Face2Face. As a product developed by AI technology, the rise of deepfake is possible today because of the emergence and progress of Generative Adversarial Networks (GANs)[2].

Figure 1: Top: Fake Videos; Bottom: True Videos.

As a result, the concept of deepfake has attracted much more attention in last several years. At the beginning of 2019, a fake video about President Trump was circulated on Facebook. In the video, “Trump” criticized Belgium's position on climate change, which is not Trump himself[3], such as Figure 1. Previously, the German research team also used this technology to produce fake videos of heads of state including Putin and Bush[4]. The U.S. Congress recently held a special hearing on the impact of deep forgery technology, which shows that deepfake technology has set off a great deal of attentions in the US election year. In Malaysia on Jan 17, 2020, The High Court of Sabah and Sarawak claims it would soon rely on AI for sentencing partly, and other forays into exploiting technology for the judiciary[5].

Technological advances have made it very easy to manipulate videos and images, and if this trend continues, evidence of photos and videos must be examined before they can be brought to courts[6]. Deepfakes are used to revenge, produce pornographic pictures of celebrities, or blackmail a person[7]. Video and photo manipulation and production have become common, mainly due to technological advances, especially in the field of machine and deep learning[8].

It's easy to generate fake information, but it's difficult to correct records and combat deepfakes[9]. Only by totally understanding the technology and principles of deepfake, we can better combat the phenomenon of deepfake.

The structure of the paper is as follows: after the introduction, the research describes the deepfake (GANs) technology. Then, applications of the deepfake technology are put forward from many aspects. We review the relevant technologies of deepfake detection from four aspects, including deepfake detection technology, multimedia forensics, anti-counterfeiting and convolution neural network. Finally, this research puts forward the insights, limitations and recommendations for future research.

2. DEEPFAKE METHODS AND APPLICATION

The term “deepfake” consists of "deep learning" and "fake". Deepfakes are the data of images, videos and voices digitally manufactured to describe things that never really happened to people.

Deepfake is the product of Generative Adversarial Networks (GANs). There are many studies related to GANs[10]–[15] such as CSGAN. The number is shown in Figure 2 (Papers on ‘GANs’ from https://scholar.google.com/, on April10, 2020).

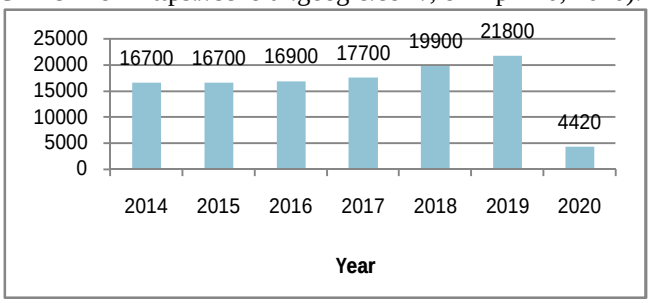

Figure 2: Number of Papers GANs Published Since 2014.

In the GANs algorithm, there are two neural networks: generator, which automatically generates samples to simulate the data in the database; and recognizer, which evaluates the authenticity of the data generated by the generator[2]. Both of them produce large-scale and high-precision output in mutual game learning. With the growing maturity and complexity of GANs, image, sound and video can be forged or synthesized automatically, and one can hardly distinguish the true from the false[10].

In Goodfellow’s paper, they proposed data is generated in the following way:

$$\lim_{\sigma \to 0} \nabla_x \, E_{\epsilon \sim N(0,\sigma^2 I)} f(x + \epsilon) = \nabla_x f(x)$$

The function of optimization of multilayer perceptrons is:

$$\min_G \max_D V(D,G) = E_{x \sim p_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} \left[\log \left(1 - D(G(z)) \right) \right]$$

Among them, D (x) represents the probability that D judges x from real data, D (G (z)) represents the probability that D judges G (z) as real data, Z represents random noise, G (z) is the probability of generating data, that is, the probability that D judges wrong, and 1-D (G (z)) is the probability of D judges right; because for D, the probability of other judges right needs to be maximized, while for G, the objective function needs to be minimized. The process is shown in Figure 3.

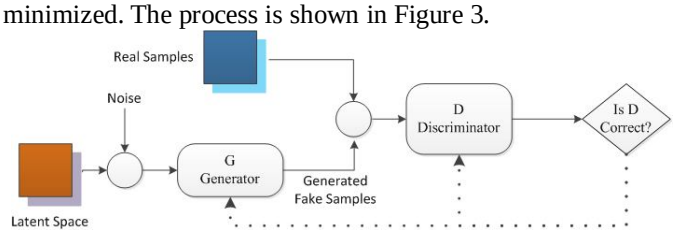

Figure 3: Architecture of GAN.

Experiments prove the huge potential of the framework, but there are the following problems:

1. Synchronization of G and D training;
2. Data generation out of control;
3. Generate some junk images or those ones similar to real data.

Since GANs waspropose in 2014[10], many models have been derived. These models of GANs are classified and shown in Table 1.

Table 1: Classification of GANs Derived Models

Type	GANs Models
GANs based on the improvement of loss function	f-GANs, Least-Square GANs, Loss-sensitive GAN, LSGANs , WGAN, WGAN-GP, WGAN-LP, DRAGAN, BEGAN, Fisher GAN, EBGANs, etc.
GANs based on Model Application Improvement	GANs based on network architecture improvement: StyleGANs, CGAN, DCGAN, InfoGAN, StackGAN, ALGAN,ASGAN, CycleGAN, DualGAN, etc.
	GANs based on encoder improvement: ProGANs, BEGAN, VAE-GAN, BiGAN, tDCGAN, Adversarial Autoencoders, etc.
	Other improved GANs: LAPGAN, MGAN, 3D-GAN, SRGAN, ESRGAN, etc.

2.1 GANs Based on the Improvement of Loss Function

Nowozin extended the variable hash estimation framework[16], proposed the f-GANs model[17], which is called Variable Divergence Minimization (VDM), and proved that the generative adversarial training is a exception of the VDM framework. The experiment shows that it has the wide application of GANs. However, the samples on the true side of the decision boundary will be classified as true. Even if they are false samples, the recognizer will still judge them as true, which leads to the gradient dispersion problem, making the quality of the pictures generated by traditional GANs not high and the training process unstable. To solve the problem, Mao proposed a least square generation countermeasure network (LSGANs)[18], which solved the problem of gradient disappearance in GANs, but could not solve the problem of how to better measure the different distance between the manufactured data and the real data.

Therefore, Arjovsky proposed that WGAN abandoned the definition of JS difference of traditional GANs, and the Earch Mover distance (EM distance) was used to calculate the distance between two distributions[19], and the evaluation problems of model collapse, training instability and model generation of GANs are solved. However, the weight clipping value in this method is not well determined, which makes it impossible to generate poor samples or convergence sometimes. In view of the shortcomings of WGAN, Gulrajani proposed an improved version of WGAN-GP, GWAN-GP[20], by adding a gradient penalty term to the discriminant function, linking the parameters with the constraints to reach the Lipschitz limit condition, so as to solve the problem of WGAN. However, WGAN-GP can't ensure the modulus of each value of the gradient is equal to or less than 1 for the region where the modulus of the gradient is greater than 1, causing the cost of this method to be very high.

GRAGAN[21]regards the alternating gradient updating process as a register minimization for training to achieve the Nash equilibrium, and proves that the model training can gradually converge under the conditions of no parameter limitation and no need for the discriminator to be in the optimal state in each step. Similarly, Berthelot proposed a new balance strengthening method – BEGAN[22], which combines the loss of EM distance to train the GANs based on the automatic encoder. Began discards the gap between the actual and generated distributions, and instead achieves the purpose of discrimination by estimating the similarity between the distribution errors of the distribution. However, it is mainly used in the image field, with general effect on high resolution images. Thus, Li added a noise removal loss function in the discriminator[23], but the effect was not more than WGAN-GP. Mroueh used the concept of singular value decomposition to embed the distribution into the finite dimension feature space[24], and match it according to its mean value and covariance feature statistics. They match the mean value feature and second-order matrix feature at the same time to maximize the potential covariance difference between the real and fake data distributions, so as to improve the training effect. In[25], they set up a framework of integral probability matrix FisherGAN to train the GANs, so that the model can be trained stably, but the constraints on the integral probability metrics (IPM) are strong and lack certain flexibility. Miyato put forward spectrum normalization GANs[26], which makes the regularization more restricted, but at the same time sacrifices the convergence speed of the model.

2.2 GANs Based on the Improvement of the Model

Radford proposed an architecture deep convolutional GANs[27], which combines the deep convolution neural network (DCGANs) of supervised learning with the GANs of unsupervised learning. DCGANs can learn a series of features no matter the discriminant model or the generating model, or the single object or the global scene of the image. During the period of training process and generating results, stability and quality have been greatly improved. Salimans put forward Improved-DCGANs in [28]. According to the training process of DCGANs, five different enhancement methods, such as small batch discrimination, historical average, single side label smoothing, were used to make the training proceed in the direction of convergence.

In view of the problem that GANs do not need to be modeled in advance, which leads to the model being too free and uncontrollable, Mirza and Osindero proposed a model with constraints on GANs, which is called CGANs for short[29]. This improvement is simple but very effective and is widely used in ([30];[31]; [32]), but it still cannot produce high-quality pictures. Zhang proposed StackGAN[33] based on CGANs to generate high-quality pictures. Since then, Zhang further proposed StackGAN++[34], to improve the model ability of processing complex text, but still unable to deal with very complex text. Johnson proposed an end-to-end method to produce images[35], which can produce complex images with many recognizable objects. Qian proposed to introduce AttentionNet into traditional GANs[36]to generate attention map, so that the network can quickly and accurately locate the focus area in the image. But with the increase of convolution size, the computational efficiency will be lost. Zhang proposed to add self attention mechanism[37] to GANs so that generators and recognizers can automatically learn important objects in images and form SAGAN, but there is still much room to be improved. Based on Sagan, BigGAN[38] applies orthogonal regularization to generator and uses hinge loss as GANs objective function, which greatly improves the fidelity and precision of generated image.

In addition, Ghosh used multi-agent GANs to improve the collapse of traditional GANs model[39]; Zhao introduced the definition of energy, and proposed the energy-based Generative Adversarial network (EBGAN)[40], resulting in higher image

resolution; Guo Jun Qi proposed the loss sensitive generation network[41] by further using the Lipschitz regularization condition of true data density to regularize its loss function, and it can improve the generalization ability of GANs; Litook the maximum mean difference as the loss function[42], and combined with the automatic encoder to generate a better model; Tolstikhn used the self-improvement training mechanism, through a fusion model, to better adapt to the data distribution, and overcome the GANs training crash[43].

2.3 Applications of GANs and Deepfakes

As a generation model, GANs can be applied to image, audio, video and other data, as well as image processing, video generation, communication protection and other tasks.

Figure 4: Images Generated by ProGAN.

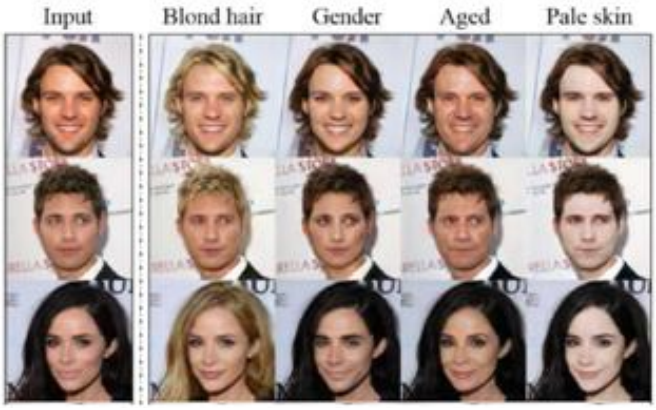

Figure 5: Images Generated by StarGAN.

In the field of images, Karras proposed ProGAN, which can generate very realistic faces [44], such as Figure 4, can greatly stabilizes and raises the training speed. Choi proposed StarGAN implements photo editing [45], mainly for face attributes. It can modify some attributes of faces, such as Figure 5, including hair color, expression, gender, age change, etc.

Cao used a 3D to assist the dual generative adversarial network (AD-GAN)[46] to precisely rotate a face image to any angle. To improve image resolution, Ledig proposed a super-resolution generation countermeasure network (SRGAN)[47], but the texture information of the generated image is noisy and not realistic. Wang proposed an enhanced resolution generation countermeasure network (ESRGAN)[48] to obtain more realistic and natural pictures (Figure 6).

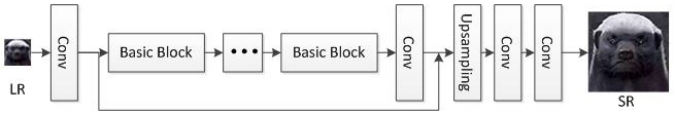

Figure 6: Architecture of ESRGAN.

In the field of information security, Triastcyn proposed a method to generate artificial data sets[49], which can retain the statistical characteristics of real data and provide differential privacy protection. Jones proposed a differential privacy assisted classification generation countermeasure network combining GANs and differential privacy to generate medical clinical data[50]. Frigerio proposed a data publishing framework to protect privacy through the definition of differential privacy[51], to ensure the protection of user personality while publishing new open data. Abadi used GANs to replace the communication parties and adversaries in the traditional symmetric encryption system with neural networks to protect the communication process [52]. Coutinho improved Abadi's model by using the concept of selective plaintext attack. Gomez proposed CipherGAN to crack some classical codes [53]. Hitaj proposed a new method, PassGAN[54], to enhance password decoding by using GANs.

Now, many deepfakes also use GANs technology. It can revive a dead man. The face changing software, called Zao, can provide that users only need to upload a photo in the app to replace their face with the characters in the short video, which can be almost fake, and now there is a lot of controversy about it. In January 2018, someone launched FakeApp, a windows program, which allows users to easily make their own face changing videos without any AI knowledge. An app called DeepNude can automatically generate naked photos as long as a female image is input. As a result of wide spread, it has caused serious adverse consequences, and the developer finally took the app off the shelf (Figure 7).

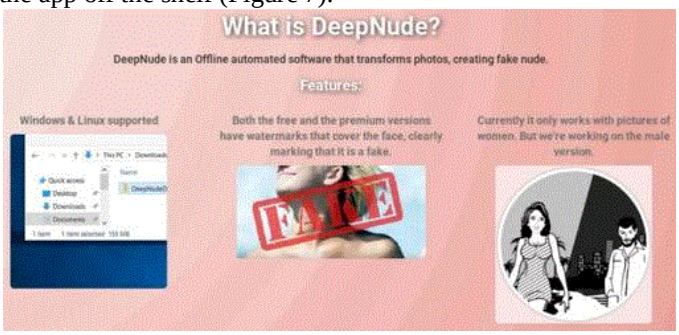

Figure 7: DeepNudeApp.

In this chapter, we mainly classify GANs from loss function and models, list the main GANs models, and analyze their advantages and disadvantages. Then describe some typical applications of GANs. All of these will be helpful for detecting deepfakes.

3. DEEPFAKE DETECTION TECHNOLOGY

As mentioned above, many fields need deepfake technology. Despite its benefits, like any other technologies, it

risks abuse. Deepfake has had a huge impact in a lot of fields. The number of papers related to deepfake is shown in Figure 8 (Papers on 'deepfake' from https://scholar.google.com/, on April 10, 2020).

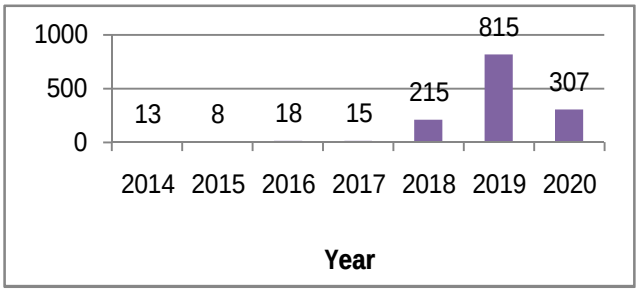

Figure 8: Number of Papers Deepfake Detection Published Since 2014.

Due to the influence of fake news, it has become a common research focus. These far-reaching implications are discussed in the following sections.

3.1 Deepfake Detection

As we all know, deepfake algorithms can replace a person's face with another face in videos or images. Although the algorithm creates a trusted changed image or video by using artificial intelligence, it still cannot generate small details such as blinking. Yuezun Li proposed a method, Long-term Recurrent Convolutional Networks (LRCN)[55], which is based on a deep learning model to combine CNNs and Recurrent Neural Networks (RNNs) to capture the phenomenology and time rule in eye-blinking process, such as Figure 9.

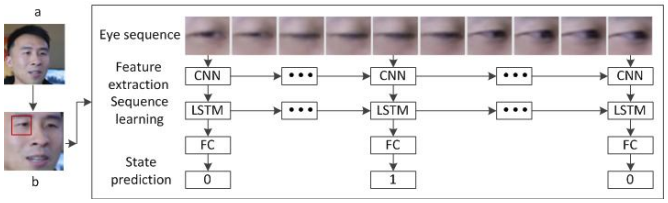

Figure 9: LRCN. (a) is the original sequence. (b) is the one after aligning face.

It shows encouraging performance in detecting the deepfake video, but the current method only uses the shortage of blink as the detection prompt, and should consider that faster blink may also be a sign of tampering, and complex forgers can still produce blink effect through post-processing. The research work[56] has described that one person usually blinks in2 to 10 seconds, and a blink takes about 1/10 to 1/4 of one second. This regularity can detect whether a video or an image is fake. If there is no correct image of a blinking person, the deepfake algorithm will produce a face image that does not blink as expected. It is emphasized that when calculating the general rate of blinking and comparing it with reality, the flicker rate of deepfake video is significantly lower. Therefore, the fake videos can be detected easily. Eye color differences can also be used to detect fakes. To do this, computer vision is used to extract the color of each eye[57]. Another way of deepfake detection is to find the details and reflections of missing teeth areas. To achieve them, the face boundary mark is cropped and adjusted to a height of 256 pixels. Then the teeth are segmented by transforming the image. Then, bright and dark clusters are collected. The bright clusters are regarded with being real and belong to the original teeth, such as Figure 10[57]. If the threshold value of mouth pixels is less than 1%, the samples will be rejected. This makes it easy to detect deepfakes.

Figure 10: Detecting Deepfake Eyes and Teeth.

Majumdar proposed a Partial Face Tampering Detection (PFTD) network[58]to detect a proposed attack. It detects the differences by mixing the high-frequency and original information of the input images, so as to capture the inconsistency among them, and the network exceeds the existing baseline of deep neural network in the performance of detecting the tampered images. But a better algorithm needs to be developed to detect the tampered area. Xu rechecked the artifacts caused by the up-sampling module of GANs, and proposed the way of signal processing and analysis. It is a new classifier design method based on spectrum input[59]. However, the proposed GAN simulator and other processing modules need to be expanded. HuyH. Nguyen proposed a capsule network [60], which can detect many kinds of attacks with fewer parameters, such as replayed and deepfake videos, such as Figure 11, but It lacks time series as input.

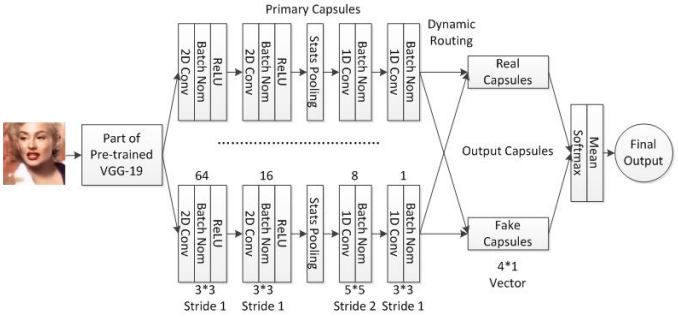

Figure 11: Architecture of Capsule-Forensics.

Yu proposed a deep learning method, that deepfake videos can be differentiated effectively from true ones[61] and there are better robustness and accuracy because of the limitations of computing time and resources, But it has not been extended to other parts of the human body except the face, and the generalization ability needs to be improved. Li proposed a novel image representation called face X-ray[62] for detecting forgery in face images. Their experiments demonstrate that face X-ray significantly improves the generalization ability through a thorough analysis. The framework achieves remarkably high detection accuracy on unseen face forgeries, as well as the ability to predict face X-rays reliably and faithfully on all kinds of recent popular face manipulations. Chih-Chung Hsu

proposed a fake face image detector[63]based on a Common Fake Feature Network (CFFN), which consists of improved DenseNet backbone network and the Siamese network architecture. The CFFN studies cross-layer characteristics and can be used to increase performance. If the Siamese network structure and detecting objects are combined, a better effect will be gotten. Jeon proposed FDFtNet[64], a novel detector of fake images and compared with the previous methods, it has 97.02% accuracy. They provide a powerful fine-tuning classifier based on neural network, which only needs a little bit of data to fine tune, and can be easily used together with other CNN architectures.

3.2 Multimedia Forensic

Nhu-tai previously proposed that anti-forensic techniques often focus on the analysis of specific correlated cues or patterns at the stages of the digital image creation or manipulation process[65]. However they built a novel deep learning network to extract face characters and obtained good results from the contest validation data, such as Figure 12.

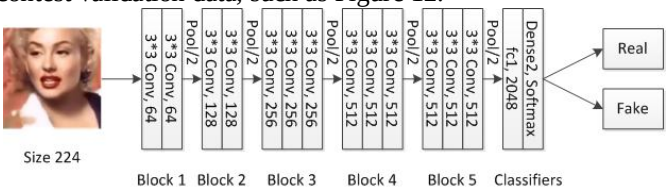

Figure 12: Architecture of Detecting Forensics.

Multimedia forensics rely on every stage of image history and uses this data to check information and determine if data or functionality has been changed. Andrea proposed to apply Photo Response Non-Uniformity (PRNU) analysis to deepfakes to evaluate the accuracy and ease of this method in deepfakes detection[66]. Haya R. Hasan proposed a solution with ethereum smart contract, so that it can track the origin and history of digital content to its original source, although the digital content has been copied many times[67] . It is still needed that distributed application and pluggable components will be developed to be used. Bourquard proposed a new way to verify whether clips of videos or images have been distorted[68], based on the proposed DIF methodology, but It also lacks the combination of data security, auxiliary channel and image forensics. Andreas proposed that even in the case of intense compression (Figure 13), deepfake detection can be used to get unprecedented accuracy, and it is significantly better than human observers [69].

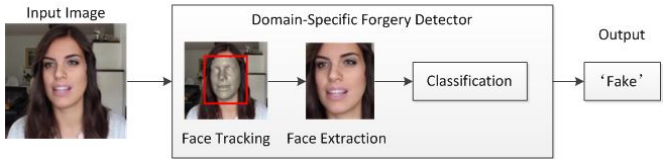

Figure 13: Pipeline of Detecting the Forgery Face.

Mengnan proposed a maneuverable fake detection method called Locality-aware AutoEncoder (LAE), which relies on correct evidence for prediction to improve the accuracy of generalization. Experimental results, based on three tasks of fake detection, show that LAE can gain high generalization accuracy by other manipulation methods [70], However, there is still a general gap, which needs to be further narrowed. Li proposed a method of pixel level analysis of face forensics using segmentation to supplement the current classification methods [71]. By redefining a problem as a pixel level task, it can evaluate lots of architectures and create a strong new benchmark for the problem.

3.3 Anti-Counterfeiting

Gardiner believed that by recognizing hidden traces, watermarks can easily identify the changed digital source [72]. By adding a watermark, it can easily be identified whether contents have been edited. Watermarks are attached when processing contents. To some extents, these traces are visible, so even if contents are shared in internet, modified elements may be accompanied by materials, which will make the recipients of fake contents alert. Cozzolino proposed a neural network to enhance the correlation traces of models hidden in video, and extracted a camera fingerprint, called video noise suppression[73]. The method based on video noise reduction in this network performed well in the main forensic tasks such as camera model recognition, video forgery and positioning, and does not need prior knowledge of specific operations or any form of fine-tuning (Figure 14). But in the process of extraction and de-noising, time direction is also used to improve the extracted noise images from videos.

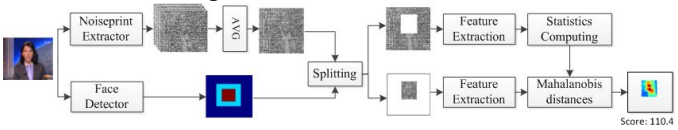

Figure 14: Block Scheme for Detecting Deepfake.

3.4 Convolutional Neural Networks

Compared with human, Convolutional Neural Networks (CNNs) and other similar methods follow the principle of machine learning and possess the ability to detect deepfake contents through powerful functions of image analysis[74]. The AI algorithms have the ability to reside information on the sharing platforms. They are running in the background, monitoring the new contents in real time, and detecting if the contents are real or fake. As a result, the technology allows users to be warned in time or to delete false contents before dissemination.

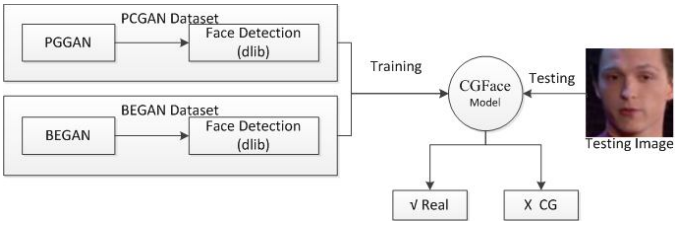

Figure 15: Computer Generated Face (CGFace).

Dang proposed a customized deep learning model, Computer Generated Face (CGFace) [75], which is specially designed for the tasks of detecting computer-generated, and implemented two most advanced methods to generate facial images with computers, in order to create two vast data sets (Figure 15), but it is worth studying other hidden features in fake images, and the combination of various functions will improve the performance of the model. Amerini proposed to use optical flow field to develop possible interframe differences[76].Then, the cue is used as the learning function of CNN classifier. The preliminary results highlight promising performance, which is obtained on the face forensics++ dataset, but it can still be improved in inconsistencies on the latest frame method and the temporal axis. Tarasiou proposed a compact architecture to detect modified face images, and it is based on full convolution neural network (FCN) [77]. The architecture can use Perceptual Generative Adversarial Networks (PGAN) and StyleGAN methods to detect fully generated images. Because of the difficulty in segmentation task, more image segmentation architecture is needed to improve the result.

4. CONCLUSION AND FUTURE RESEARCH

The progress of artificial intelligence technology improves the data tampering technology, and sometimes the data can be used as evidence. Maras reported that deepfake technology can search different websites to find images of the target people[78]. Gardiner suggested that deepfake technologies can be widely used in multimedia industries, such as films and televisions, as well as large-scale social applications[72]. However, this technology is used unethically and illegally. Deepfakes can reduce the authenticity of evidence, and even affect people's lives.

As a result of the heavy use of deepfake, which caused a lot of criticism from netizens, Reddit closed the deepfakes forum and updated the rules of the entire network, stating that websites that involve involuntary pornography and underage hints will be scrutinized. Social platforms like Twitter, Pornhub (pornographic sites), Gfycat (GIF dynamic image platform), etc. have also resisted this. Facebook has also established a model to detect fake images or videos.

Zannettou listed some subjects related to deepfake[79], including government, companies, criminals and individuals, and these people's motivations may include maliciously harming others, manipulating public opinion, creating celebrity scandals or harassing someone, distributing negative false information to the public, or, as described by MacKenzie and Bhatt, purely fun and entertainment[80]. There are also individuals and organizations that produce and support deepfakes for legitimate uses, such as paid work for music videos. Despite the many shortcomings of deepfake, there are some legitimate applications to consider in education media, digital communications, gaming and entertainments, healthcare, materials science, and other business areas.

Now there are several methods to counter deepfakes: 1) supervision and legislation, 2) company policy, 3) publicity, 4) training and education, and 5) anti-deepfake technology[81].Everyone should improve the awareness of the threat of deepfake technology to society, and governments should also legislate to prohibit the use of deepfake technology for immoral business, political or anti social purposes[82]–[85]. Deepfake detection is far from being solved. In fact, using deepfake and other technologies to reliably detect their pursuit is often endless. This is because the essence of the problem is "cat and mouse".

For the emerging GANs models, deepfake detection technology usually lags one step. Wang proposed that there are some common defects in the image generated by CNNs, which makes the synthesized image always have some flaws [86]. They detected the deepfake images generated by 11 common GANs models and achieved tremendous results. Based on the method, we will analyze the spectrum of datasets and try to find a more generalized and effective method to detect deepfakes generated by as many GANs as possible.

This paper reviews the models related to GANs and deepfakes, and their applications in many fields in recent years. These technologies have been widely used. At the same time, they promote and develop each other, and make positive contributions to society. Just as a knife has two sides, it can be used both positively and negatively. Any technology can be exploited by evil motives. At present, relying on AI to detect fake contents is still the most feasible solution. But there are many important ways to explore. For instance: how to conduct manual verification skillfully in technical testing? Should these technologies be open source to some extent before being controlled? How can relevant policies not hinder the beneficial development of technology?

In the near future, with the maturity and improvement of new models such as capsule network[87], spiking neural network[88] and the classifier model based on spectrum instead of pixel[86], which have made significant progresses, there will be great development potential in deepfakes.

ACKNOWLEDGMENT

The work of this paper is supported by the Zhejiang Provincial Natural Science Foundation of China(LY20G010007).

REFERENCES

1. L. Borges, B. Martins, and P. Calado, **Combining similarity features and deep representation learning for stance detection in the context of checking fake news**, *J. Data Inf. Qual.*, vol. 11, no. 3, pp. 1–26, 2019. https://doi.org/10.1145/3287763
2. J. Gui, Z. Sun, Y. Wen, D. Tao, and J. Ye, **A Review on**

Generative Adversarial Networks: Algorithms, Theory, and Applications, *arXiv Prepr.*, vol. 14, no. 8, pp. 1–28, 2020.

3. H. Von Der Burchard, **Belgian socialist party circulates deepfake Donald Trump video**, *www.politico.eu*, 2018. https://www.politico.eu/article/spa-donald-trumpbelgium-paris-climate-agreement-belgian-socialist-party-circulates-deep-fake-trump-video/.
4. S. A. Topol, **What Does Putin Really Want**, *www.nytimes.com*, 2019. https://www.nytimes.com/2019/06/25/magazine/russia-united-states-world-politics.html.
5. **Chief Judge Sabah Sarawak High Courts to Use AI to Propose Punishments**, *www.malaymail.com*, 2020. https://www.malaymail.com/news/malaysia/2020/01/17/chief-judge-sabah-sarawak-high-courts-to-use-ai-to-propose-punishments/1828985#.XiJRCJH3y04.facebook.
6. R. Chesney and D. K. Citron, **Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security**, *SSRN Electron. J.*, vol. 107, p. 1753, 2019. https://doi.org/10.2139/ssrn.3213954
7. D. Harris, **Deepfakes: False Pornography Is Here and the Law Cannot Protect You**, *Duke Law Technol. Rev.*, vol. 17, p. 99, 2019.
8. D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, **MesoNet: A compact facial video forgery detection network**, *2018 IEEE Int. Work. Inf. Forensics Secur.*, pp. 1–7, 2018. https://doi.org/10.1109/WIFS.2018.8630761
9. A. De keersmaecker, J., & Roets, **"Fake news": Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions**, *Intelligence*, vol. 65, pp. 107–110, 2017. https://doi.org/10.1016/j.intell.2017.10.005
10. I. J. Goodfellow *et al.*, **Generative adversarial nets**, in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.
11. M. Kocaoglu, C. Snyder, and A. G. Dimakis, **CausalGAN : Learning Causal Implicit Generative Models with Adversarial Training**, *arXiv Prepr.*, 2017.
12. Jay Robert, B. Del Rosario, **Development of a Face Recognition System Using Deep Convolutional NeuralNetwork in a Multi-view Vision Environment**,*International Journal of Advanced Trends in Computer Science and Engineering*, Volume 8, No.3, pp.369-374, 2019. https://doi.org/10.30534/ijatcse/2019/06832019
13. J. Zhao, J. Li, Y. Cheng, T. Sim, S. Yan, and J. Feng, **Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing**, *Proc. 26th ACM Int. Conf. Multimed.*, pp. 792–800, 2018. https://doi.org/10.1145/3240508.3240509
14. A. Jahanian, L. Chai, and P. Isola, **On the "steerability" of** generative **adversarial networks**, *arXiv Prepr.*, 2019.
15. Aaron Don M. Africa, Ara Jyllian A. Abello, Zendrel G. Gacuya, Isaiah Kyle A. Naco,Victor Antonio R. Valdes, **Face Recognition Using MATLAB**, *International Journal of Advanced Trends in Computer Science and Engineering*, Volume 8, No.4, pp.1110-1116, 2019. https://doi.org/10.30534/ijatcse/2019/17842019
16. X. Nguyen, M. J. Wainwright, and M. I. Jordan, **Estimating divergence functionals and the likelihood ratio by convex risk minimization**, *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5847–5861, 2010. https://doi.org/10.1109/TIT.2010.2068870
17. S. Nowozin, B. Cseke, and R. Tomioka, **f-gan: Training generative neural samplers using variational divergence minimization**, *Adv. Neural Inf. Process. Syst.*, pp. 271–279, 2016.
18. X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, **Least Squares Generative Adversarial Networks**, in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2794–2802. https://doi.org/10.1109/ICCV.2017.304
19. M. Arjovsky, S. Chintala, and L. Bottou, **Wasserstein generative adversarial networks**, in *34th International Conference on Machine Learning, ICML 2017*, 2017.
20. I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, **Improved training of wasserstein GANs**, in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.
21. N. Kodali, J. Abernethy, J. Hays, and Z. Kira, **On Convergence and Stability of GANs**, *arXiv Prepr.*, 2017.
22. D. Berthelot, T. Schumm, and L. Metz, **BEGAN: Boundary Equilibrium Generative Adversarial Networks**, *IEEE Access*, 2017.
23. Y. Li, N. Xiao, and W. Ouyang, **Improved boundary equilibrium generative adversarial networks**, *IEEE Access*, vol. 6, pp. 11342–11348, 2018. https://doi.org/10.1109/ACCESS.2018.2804278
24. Y. Mroueh, T. Sercu, and V. Goel, **McGan: Mean and covariance feature matching GAN**, *arXiv Prepr.*, pp. 2527–2535, 2017.
25. Y. Mroueh and T. Sercu, **Fisher GAN**, *Adv. Neural Inf. Process. Syst.*, pp. 2510–2520, 2017.
26. T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, **Spectral normalization for generative adversarial networks**, *arXiv Prepr.*, 2018.
27. A. Radford, L. Metz, and S. Chintala, **Unsupervised representation learning with deep convolutional generative adversarial networks**, *arXiv Prepr.*, 2015.
28. T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, **Improved techniques for training GANs**, *Adv. Neural Inf. Process. Syst.*, pp. 2234–2242, 2016.
29. J. Gauthier, **Conditional generative adversarial nets for convolutional face generation**, *Cl. Proj. Stanford CS231N Convolutional Neural Networks Vis. Recognition, Winter semester*, vol. 2014, no. 5, p. 2, 2014.
30. A. Odena, C. Olah, and J. Shlens, **Conditional image synthesis with auxiliary classifier gans**, *34th Int. Conf. Mach. Learn. ICML 2017*, vol. 70, pp. 2642–2651, 2017.
31. P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, **Image-to-image translation with conditional adversarial networks**, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

https://doi.org/10.1109/CVPR.2017.632

32. Hamad Almohamedh, Sultan Almotairi, **Facial Emotion Recognition Using Eigenface and Feature Optimization**, *International Journal of Advanced Trends in Computer Science and Engineering,*Volume 8, No.4, pp.1181-1185, 2019. https://doi.org/10.30534/ijatcse/2019/28842019
33. H. Zhang *et al.*, **StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks**, *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 5908–5916, 2017.
34. H. Zhang *et al.*, **StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks**, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1947–1962, 2019. https://doi.org/10.1109/TPAMI.2018.2856256
35. J. Johnson, A. Gupta, and L. Fei-Fei, **Image Generation from Scene Graphs**, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1219–1228.
36. R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu, **Attentive Generative Adversarial Network for Raindrop Removal from A Single Image**, *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 2482–2491, 2018. https://doi.org/10.1109/CVPR.2018.00263
37. H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, **Self-attention generative adversarial networks**, *arXiv Prepr.*, 2018.
38. A. Brock, J. Donahue, and K. Simonyan, **Large scale GaN training for high fidelity natural image synthesis**, *arXiv Prepr.*, 2018.
39. A. Ghosh, V. Kulharia, V. Namboodiri, P. H. S. Torr, and P. K. Dokania, **Multi-agent Diverse Generative Adversarial Networks**, in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8513–8521.
40. J. Zhao, M. Mathieu, and Y. LeCun, **Energy-based generative adversarial networks**, *arXiv Prepr.*, 2016.
41. G. J. Qi, **Loss-Sensitive Generative Adversarial Networks on Lipschitz Densities**, *Int. J. Comput. Vis.*, pp. 1–23, 2019.
42. Y. Li, K. Swersky, and R. Zemel, **Generative moment matching networks**, in *International Conference on Machine Learning*, 2015, pp. 1718–1727.
43. I. Tolstikhin, S. Gelly, O. Bousquet, C. J. Simon-Gabriel, and B. Schölkopf, **AdaGAN: Boosting generative models**, in *Advances in Neural Information Processing Systems*, 2017, pp. 5424–5433.
44. T. Karras, T. Aila, S. Laine, and J. Lehtinen, **Progressive growing of GANs for improved quality, stability, and variation**, *arXiv Prepr.*, 2017.
45. Y. Choi, M. Choi, M. Kim, J. W. Ha, S. Kim, and J. Choo, **StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation**, *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, pp. 8789–8797, 2018. https://doi.org/10.1109/CVPR.2018.00916
46. J. Cao, Y. Hu, B. Yu, R. He, and Z. Sun, **3D Aided Duet GANs for Multi-View Face Image Synthesis**, *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 8, pp. 2028–2042, 2019.
47. C. Ledig *et al.*, **Photo-realistic single image super-resolution using a generative adversarial network**, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690. https://doi.org/10.1109/CVPR.2017.19
48. X. Wang *et al.*, **ESRGAN: Enhanced super-resolution generative adversarial networks**, in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
49. B. Triastcyn, A., & Faltings, **Generating Differentially Private Datasets Using Gans**, 2018.
50. B. K. Beaulieu-Jones *et al.*, **Privacy-preserving generative deep neural networks support clinical data sharing**, *Circ. Cardiovasc. Qual. Outcomes*, vol. 12, no. 7, pp. 1–10, 2019.
51. L. Frigerio, A. S. de Oliveira, L. Gomez, and P. Duverger, **Differentially private generative adversarial networks for time series, continuous, and discrete open data**, *IFIP Int. Conf. ICT Syst. Secur. Priv. Prot.*, pp. 151–164, 2019.
52. M. Coutinho, R. de O. Albuquerque, F. Borges, L. J. G. Villalba, and T. H. Kim, **Learning perfectly secure cryptography to protect communications with adversarial neural cryptography**, *Sensors*, vol. 18, no. 5, p. 1306, 2018.
53. A. N. Gomez, S. Huang, I. Zhang, B. M. Li, M. Osama, and L. Kaiser, **Unsupervised Cipher Cracking Using Discrete GANs**, *arXiv Prepr.*, 2018.
54. B. Hitaj, P. Gasti, G. Ateniese, and F. Perez-Cruz, **PassGAN: A deep learning approach for password guessing**, *Int. Conf. Appl. Cryptogr. Netw. Secur.*, pp. 217–237, 2019. https://doi.org/10.1007/978-3-030-21568-2_11
55. Y. Li, M.-C. Chang, H. Farid, and S. Lyu, **In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking**, *2018 IEEE Int. Work. Inf. Forensics Secur.*, pp. 1–7, 2018.
56. R. Chawla, **Deepfakes: How a pervert shook the world**, *Int. J. Adv. Res. Dev.*, vol. 4, no. 6, pp. 4–8, 2019.
57. F. Matern, C. Riess, and M. Stamminger, **Exploiting visual artifacts to expose deepfakes and face manipulations**, *Proc. - 2019 IEEE Winter Conf. Appl. Comput. Vis. Work. WACVW 2019*, pp. 83–92, 2019.
58. P. Majumdar, A. Agarwal, R. Singh, and M. Vatsa, **Evading Face Recognition via Partial Tampering of Faces**, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Work.*, 2019. https://doi.org/10.1109/CVPRW.2019.00008
59. X. Zhang, S. Karaman, and S.-F. Chang, **Detecting and Simulating Artifacts in GAN Fake Images**, *arXiv Prepr.*, 2019.
60. H. H. Nguyen, J. Yamagishi, and I. Echizen, **Use of a Capsule Network to Detect Fake Images and Videos**, *arXiv Prepr.*, 2019.
61. C. Yu, C. Chang, and Y. Ti, **Detecting Deepfake-Forged Contents with Separable Convolutional Neural**

Network and Image Segmentation, *arXiv Prepr.*, 2019.
62. L. Li *et al.*, **Face X-ray for More General Face Forgery Detection**, *arXiv Prepr.*, 2019.
63. C.-C. Hsu, Y.-X. Zhuang, and C.-Y. Lee, **Deep Fake Image Detection Based on Pairwise Learning**, *Appl. Sci.*, vol. 10, no. 1, p. 370, 2020.
64. H. Jeon, Y. Bang, and S. S. Woo, **FDFtNet: Facing Off Fake Images using Fake Detection Fine-tuning Network**, *arXiv Prepr.*, 2020.
65. N. Do, I. Na, and S. Kim, **Forensics Face Detection From GANs Using Convolutional Neural Network**, 2018.
66. M. Albahar and J. Almalki, **DEEPFAKES: THREATS AND COUNTERMEASURES SYSTEMATIC REVIEW**, *J. Theor. Appl. Inf. Technol.*, vol. 97, no. 22, 2019.
67. H. R. Hasan and K. Salah, **Combating Deepfake Videos Using Blockchain and Smart Contracts**, *IEEE Access*, vol. 7, pp. 41596–41606, 2019. https://doi.org/10.1109/ACCESS.2019.2905689
68. A. Bourquard and J. Yan, **Differential Imaging Forensics**, *arXiv Prepr.*, Jun. 2019, Accessed: Mar. 16, 2020.
69. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, **FaceForensics++: Learning to Detect Manipulated Facial Images**, *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1–11, 2019.
70. M. Du, S. Pentyala, Y. Li, and X. Hu, **Towards Generalizable Forgery Detection with Locality-aware AutoEncoder**, *arXiv Prepr.*, 2019.
71. J. Li, T. Shen, W. Zhang, H. Ren, D. Zeng, and T. Mei, **Zooming into Face Forensics: A Pixel-level Analysis**, *arXiv Prepr.*, 2019.
72. N. Gardiner, **Facial re-enactment, speech synthesis and the rise of the Deepfake**, *Res. Online*, 2019.
73. D. Cozzolino, G. Poggi, and L. Verdoliva, **Extracting camera-based fingerprints for video forensics**, *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Work.*, pp. 130–137, 2019.
74. D. Guera and E. J. Delp, **Deepfake Video Detection Using Recurrent Neural Networks**, in *Proceedings of AVSS 2018 - 2018 15th IEEE International Conference on Advanced Video and Signal-Based Surveillance*, 2019, pp. 1–6.
75. L. M. Dang, S. I. Hassan, S. Im, J. Lee, S. Lee, and H. Moon, **Deep learning based computer generated face identification using convolutional neural network**, *Appl. Sci.*, vol. 8, no. 12, p. 2610, 2018. https://doi.org/10.3390/app8122610
76. I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, **Deepfake Video Detection through Optical Flow Based CNN**, in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.
77. M. Tarasiou and S. Zafeiriou, **Using Fully Convolutional Neural Networks to detect manipulated images in videos**, *arXiv Prepr.*, 2019.
78. M. H. Maras and A. Alexandrou, **Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos**, *Int. J. Evid. Proof*, vol. 23, no. 3, pp. 255–262, 2019.
79. S. Zannettou, M. Sirivianos, J. Blackburn, and N. Kourtellis, **The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans**, *J. Data Inf. Qual.*, vol. 11, no. 3, pp. 1–37, 2019. https://doi.org/10.1145/3309699
80. A. MacKenzie and I. Bhatt, **Lies, Bullshit and Fake News: Some Epistemological Concerns**, *Postdigital Sci. Educ.*, vol. 2, no. 1, pp. 9–13, 2020.
81. M. Westerlund, **The Emergence of Deepfake Technology: A Review**, *Technol. Innov. Manag. Rev.*, vol. 9, no. 11, 2019.
82. K. E. Anderson, **Getting acquainted with social networks and apps: combating fake news on social media**, *Libr. Hi Tech News*, vol. 35, no. 3, pp. 1–6, 2018.
83. Sukhada Chokkadi, Sannidhan MS, Sudeepa K B, Abhir Bhandary, **A Study on various state of the art of the Art Face Recognition System using Deep Learning Techniques**,*International Journal of Advanced Trends in Computer Science and Engineering*, Volume 8, No.4, pp.1590-1600, 2019. https://doi.org/10.30534/ijatcse/2019/84842019
84. L. Floridi, **Artificial Intelligence, Deepfakes and a Future of Ectypes**, *Philos. Technol.*, vol. 31, no. 3, pp. 317–321, 2018.
85. R. Spivak, **"DEEPFAKES": THE NEWEST WAY TO COMMIT ONE OF THE OLDEST CRIMES**, vol. 3, no. 2, pp. 339–400, 2019.
86. S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, **CNN-generated images are surprisingly easy to spot... for now**, *arXiv Prepr.*, 2019.
87. H. H. Nguyen, J. Yamagishi, and I. Echizen, **Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos**, in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2019, pp. 2307–2311. https://doi.org/10.1109/ICASSP.2019.8682602
88. S. Kim, S. Park, B. Na, and S. Yoon, **Spiking-YOLO: Spiking Neural Network for Energy-Efficient Object Detection**, no. December, 2019.