Table of Contents

DeepFakes: Detecting Forged and Synthetic Media Content Using Machine Learning

Sm Zobaed[1], Md Fazle Rabby[1], Md Istiaq Hossain[2], Ekram Hossain[1], Md sazib Hasan[3], *Asif Karim[4], and Khan Md. Hasib[5]

[1] University of Louisiana, Lafayette, LA 70503, USA
[2] Southern Utah University, Cedar City, UT 84720, USA
[3] Dixie State University, St. George, UT 84770, USA
[4] Charles Darwin University, Casuarina, NT 0810, Australia
[5] Ahsanullah University of Science Technology, Dhaka, Bangladesh
{sm.zobaed1,ekram.hossain1}@louisiana.edu, {sourav.sust.cse.10, shishir2004x, khanmdhasib.aust}@gmail.com, mdsazibhasan@dixie.edu, asif.karim@cdu.edu.au

Abstract. The rapid advancement in deep learning makes the differentiation of authentic and manipulated facial images and video clips unprecedentedly harder. The underlying technology of manipulating facial appearances through deep generative approaches, enunciated as *DeepFake* that have emerged recently by promoting a vast number of malicious face manipulation applications. Subsequently, the need of other sort of techniques that can assess the integrity of digital visual content is indisputable to reduce the impact of the creations of DeepFake. A large body of research that are performed on DeepFake creation and detection create a scope of pushing each other beyond the current status. This study presents challenges, research trends, and directions related to DeepFake creation and detection techniques by reviewing the notable research in the DeepFake domain to facilitate the development of more robust approaches that could deal with the more advance DeepFake in future.

Keywords: DeepFake Generation · DeepFake Detection · Adversarial Attack · Face Swap

1 Introduction

Because of the advances of deep learning and generative adversarial networks (GAN) [1], creation of a realistically looking face image of a target person who really does not exist or altercation of facial appearance (attributes, identity, expression) is attainable with maintaining realism. The deep learning research community roughly refers to the technology as "DeepFake" that is coined from "deep learning" and "fake". Generally, DeepFake approaches require a massive volume of image and video data to train models for generating realistic images and videos. Because of the wide availability of robust pre-trained DeepFake models, malicious DeepFake contents are generated that create negative impact on the societies.

Pre-print: will be available in Springer Nature

The potential target of DeepFake is the public figures such as celebrities, priests, and politicians whose videos and images are largely available on the internet. More specifically, DeepFake is often used to alter faces of celebrities or politicians to other bodies in pornographic contents. DeepFakes can be abused to create political or ethnic tensions between countries to fool common people to affect election, or create chaos in sports or global economy by creating fake contents.

There are numerous notable examples of DeepFake incidents that have been shared in the internet [2, 3, 4, 5]. For instance, in 2018, a video posted in Facebook showing Former President of USA, Donald Trump taunted Belgium for remaining in the Paris climate agreement [4]. By noticing the video clearly, it was determined that Trump's hair looked stranger than usual and his voice was rolled up. In 2019, a DeepFake video of Facebook owner, Mark Zuckerberg, was published on Instagram [4]. In the video, Zuckerberg's speech was altered along with his facial expression so that the viewers can easily be distracted. A recent release of an app named DeepNude raises issue since it is used to transform a person to a non-consensual pornography [6]. Similarly, a Chinese app named "Zao" got viral lately for offering face swapping with bodies of TV stars and even replace themselves into well-known movies and TV clips [7]. These forms of manipulation create a serious threat to privacy and identity, and even jeopardize personal lives. Although the evil technology is undoubtedly a severe threat to world security, it is also used in positive purposes such as updating episodes of a visual content even after the actor is dead or creating speech of mute people. However, the number of maliciously used cases DeepFake significantly outperforms the number of positively used cases.

The underlying mechanism for DeepFake creation is advanced deep learning models such as autoencoders and GAN, which have been applied widely in the computer vision research community. Due to the development of advanced deep networks and the availability of a substantial amount of training data, manipulated images and videos have turned out almost indistinguishable to human eyes and even to robust algorithms. Hence, the creation of those manipulated contents becomes simpler and takes comparatively lesser effort. This is because an identity image or small video clip of a targeted individual are sufficient for the inference tasks.

The rise of stunning DeepFake creation vividly highlights the significance of judging the genuineness of digital media content. Because of the availability of various DeepFake creation tools, almost anyone can simply create forged content these days. As a result, in the computer vision research community, the study of DeepFake has gained traction in recent years for detecting such contents [8, 9, 10, 11, 12]. In [10], Juefei-Xu *et al.* showed a distribution of DeepFake related papers in last 5 years, where 78% of the total papers published in the last two years. This increase amount of paper in the last two years vividly highlights research interest revolved around DeepFakes.

Governments and law enforcement are undertaking the spread of DeepFake creations with new policies and regulations as well. For example, US Senator named

Ben Sasse proposed a bill titled *S.3805 - Malicious deep fake prohibition act of 2018* in 2018 that introduces a new type of criminal offense because of the creation or distribution of fake digital media content that falsify realism [13]. Besides, social media platforms (*e.g.,* Twitter, Facebook) are actively taking initiatives to deal with forged, synthetic, and manipulated content on their respective platforms. For example, in Twitter, if a tweet contains manipulated media content specially, DeepFakes, Twitter has started to alert users about that by tagging with warning sign and attaching the trustworthy news article link relevant to the tweet [14]. In another example, in 2019, Facebook facilitated the development of robust DeepFake detection tools by organizing the DeepFake detection challenge (DFDC) where 2114 number of participants across the globe had participated and they generated more than 35,000 models [15].

In Figure 1, we depict the relation between number of papers that are related to DeepFake in years from 2015 to 2020. The data is collected from Google Scholar on April 2021 with the query keyword "deepfake" found in either title or full text of the papers. According to the number of related papers has increased significantly in the recent years which is an indication Deep Fakes related research or news are getting noticed a lot more in recent times.

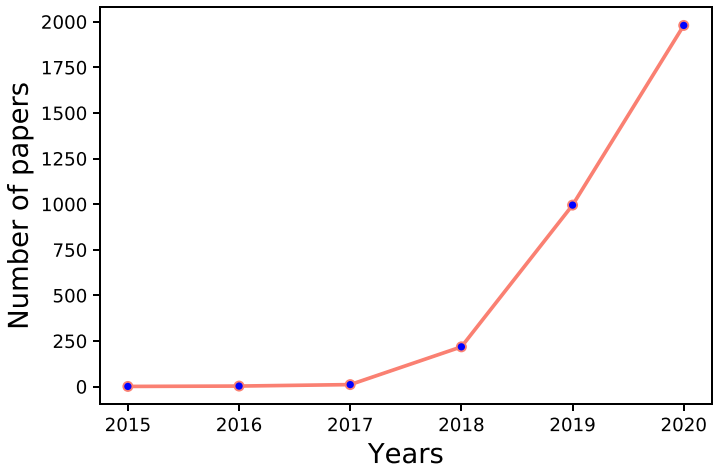

Fig. 1: Number of scholarly articles that are related to DeepFake in years from 2015 to 2020, obtained from Google Scholar on April 2021 with the query keyword "deepfake" applied to either title or full text of the papers.

This chapter presents a handful number of methods for DeepFake generation and detection in comprehensive manner. In Section 2, we discuss how deep learning is leveraged on DeepFake algorithms for creating manipulated contents. Section 3 reviews a wide set of effective methods for DeepFake detection as well as their advantages and disadvantages. In Section 4, We discuss challenges, re-

search trends, and directions on DeepFake generation and detection domains. Finally, Section 5 concludes the study.

2 DeepFake Generation

DeepFake contents get attraction because of the availability of robust and powerful set of DeepFake applications to a wide range of users. Such applications are capable to create a forged content by a few number of clicks within a few seconds. Because of their rapid popularity, a large number of researches have been performed on DeepFake generation in recent days. In this section, we discuss about various DeepFake generation approaches and the datasets.

2.1 Generation Approaches

Most of the DeepFake related works have been done leveraging deep learning techniques. From the state-of-the-art literatures, DeepFake generation through facial image manipulation can be classified into four methodological categories based on the way and extent of manipulation: complete face synthesis, identity swap, attribute manipulation, and Face Reenactment. We provide a detailed discussion in the following sections. We discuss all other methods in a separate group - "Other DeepFake Generation Methods".

2.1.1 Complete Face Synthesis Facial manipulation or editing techniques have been studied and developed considerably over the last few decades. It is also practised for generating DeepFake contents in recent days. Face synthesis generates photorealistic images of human faces that do not exist in real life. With the massive progression in Generative models, GAN, [16] in the last few years, the research community has seen a significant amount of works associated with facial manipulation. GANs have been effectively used for generating photorealistic face images. Another variation of the generative deep learning model is variational autoencoder (VAE) [17] that also shows the potentiality in creating human face image.

Initially, these adversarial model starts generating realistic fake images from random vectors. The generative model tries to generate a more realistic image and fool the discriminative model in each iteration. On the contrary, the goal of the discriminative model is to verify the generated photo is either real or fake. Radford *et al.* proposed the deep convolutional generative adversarial network (DCGAN) [18], where the concept of both GAN and Convolutional Neural Network (CNN) has been utilized together to create a nonexisting human face. It is one of the initial works after the emergence of GAN in 2014. Liu *et al.* proposed VAE based CoGAN [19]. In COCO-GAN [20], the authors proposed a conditional GAN-based image generator that is capable of synthesizing images in a parallelizable fashion. However, Glow [21], a flow-based generative model, which is different from GAN's mechanism, is proposed by Kingma e*et al.* . In

this work, the authors used invertible 1x1 convolution for generating realistic DeepFake images.

Later in 2017, Wasserstein generative adversarial networks (WGAN) [22] has been proposed. The approach used in WGAN training is more stable than the previous method. Stability in GANs training was one of the primary issues in the first few years right after GAN's invention. WGAN minimizes this instability in GAN training. However, Gulrajani *et al.* [23] showed that due to the weight clipping operation, sometimes WGAN might fail to converge, thus might generate lousy images as output. In this paper, they provide an improved weight clipping approach to address the issue in WGAN training. BEGAN [24] is another work with the aim of improving WGAN. Karras *et al.* presented Progressive Growing GAN (PGGAN) [25] in 2017 with the focus on generating high-quality images. This is one of the pioneering works on generating high-quality images. The same author proposed StyleGAN [26] in 2019 that can automatically learn the high-level attribute representation such as identity, pose to control different properties in generated images. StyleGAN2 [27], the extended version of the previous work, was presented in 2020.

2.1.2 Identity Swap Identity swap is one of the most common face manipulation research techniques associated with DeepFakes. This approach includes replacing the human face in the target content (image or video) with another face in the source content. The traditional face swap process can be performed in three phases. First, the face is required to be detected in both source, and target content that can be done with face detection [28], or object detection model [29, 30]. The research community has seen numerous defensive [31] and offensive [32] applications with object detection techniques. After face or facial attributes detection in source and target content, the eyes, nose, eyebrows, mouth is replaced and adjusted and blended in term of lighting and color to minimize the difference between source and target content. In the third step, the adjusted candidates are ranked by the calculated distance over the overlapped region. However, this traditional face swap approach has limitations in generating very realistic face images as it offers static and rigid replacement. Different DL-based approaches have become very effective in realistic face-swapping with the super-progress in the Deep learning (DL) domain.

FaceSwap [33], and CycleGAN [34] are some of the very first works in this field. In FaceSwap, two sets of encoder-decoder combinations are used. The encoder part of the architecture is responsible for composing the latent feature of a face from the input image, and then the decoder part reconstructs the face. There are two parts of the training phase. In the first phase, Each encoder-decoder combination is trained with the source image. In the second phase of the training, the decoder gets trained with the target image. After successful training, the two decoders are substituted with each other. Consequently, the original encoder paired with the decoder of the target image is capable of constructing the target image with the facial features of the source image. The

DeepFake generation (identity swap) procedure with pairs of encoder-decoder architecture is illustrated in Figure 2.

The CycleGAN [34] solved the issue of the unavailability of paired training examples for image translation. FSGAN [35] is capable of face swapping and reenactment simultaneously with face reenactment and blending. Natsume *et al.* proposed two distinct VAE based RSGAN [36] to encode the latent representation of facial attributes. The recent work, FaceShifter [37] uses a two-phase scheme for high fidelity and occlusion-based face-swapping.

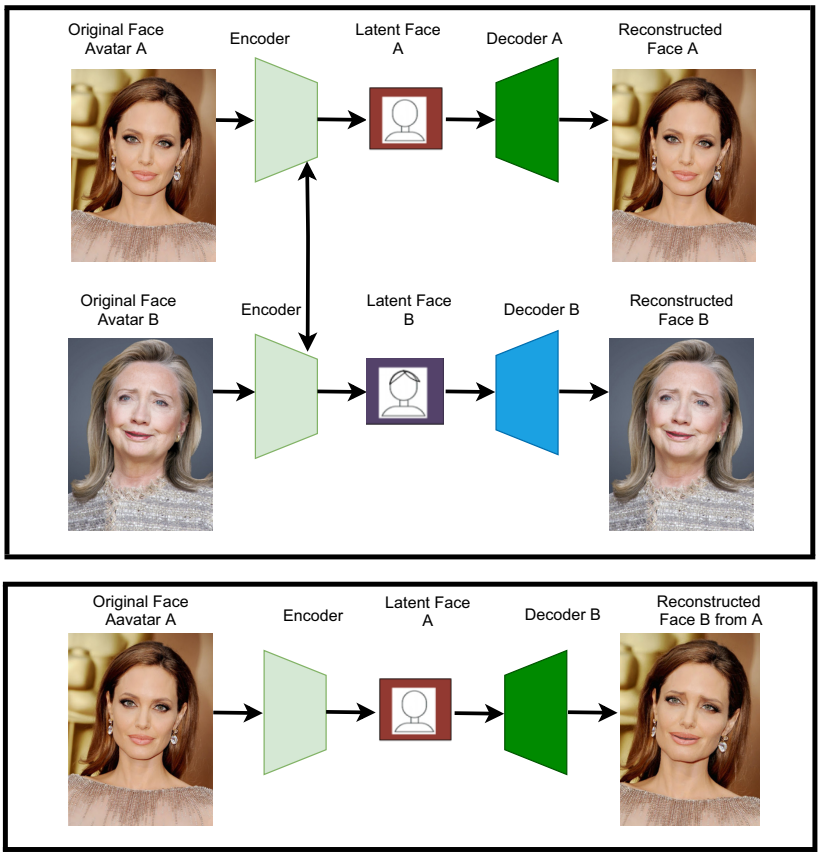

Fig. 2: A DeepFake creation model using two encoder-decoder combinations. In the training phases, the encoder-decoder set is used to learn the latent features of the input faces. While generation, two decoders are interchanged, such that latent face A is subjected to decoder B to generate the face A including the features of face B.

2.1.3 Attribute Manipulation Face attribute manipulation is a process of modifying the specific region of the face in the target image. This process is very similar to face editing in most cases. Some examples of face manipulation are changing age, gender, hair color and style, disappearing hair (bald), and creating smiling faces from neutral faces, etc. Choi *et al.* applied the GAN concept in the image-to-image translation problem by presenting StarGAN [38] in 2018. In StarGAN, There is a single generator for translating images from a domain to multiple domains. However, StarGAN has a limitation as it can only generate a specific number of expressions. For Image translation, Chen *et al.* proposed HomointerpGAN [39], where during translation intermediate region between different domains is considered. The author suggested proper methods to select paths between two sample points in latent space to change particular image attributes.

Pumarola *et al.* mitigated this issue by proposing GANimation [40] where action unit (AU) annotation based GAN conditioning method is implemented. Later, the authors presented the improved version StarGANv2 [41] that can generate images with the highest visual quality. To develope this improved version, the authors design the model with encoder-decoder architecture where a random Gaussian noise vector fed to the generator. In terms of expression synthesis and attribute manipulation, StarGANv2 outperforms other works for its high scalability. To achieve better. He *et al.* introduced encoder-decoder-based AttGAN [42] where conditioned latent representation is used for more specific facial attribute editing as well as preserving other details of the face. One of the limitations of AttGAN is unwanted blurriness in generated face images. STGAN [43] is proposed by Liu *et al.* in 2019 as an improvement of AttGAN. In this work, the difference between target and source attribute is taken into consideration for more specific face attribute editing. However, STGAN shows poor performance in multiple attribute manipulation in the face image.

2.1.4 Face Reenactment Face reenactment is a type of emerging DeepFake face manipulation technique, more precisely, which can be stated as a conditional face synthesis task for facial expression transfer. It refers to a process that replaces or transfers the facial expression of a person to another person. Face reenactment can be achieved by transferring the source actor's expression, gaze, pose, and mouth movement. Some of the most prominent works for real-time facial expression transfer have been done by Theis *et al.* [44, 45]. In Face2Face, the authors proposed methodologies to transfer the source person's facial expressions (actor), including facial gestures, head, and eye movement, to a video with another person (target person), maintaining the identity. Face2Face approach use deformation transfer between source face and target face, more specifically the mouth portion with higher priority for photo-realistic reenactment. For the tracking and reconstruction of the face identity (3D face model) of a source and target model, a commodity RGB-D sensor is used. After getting the required parameters from the 3D models, face expression is reenacted to the target face in each generated fake video frame. This approach is applicable in real-time stan-

dard RGB videos. In the same research group's successive work, FaceForensic++ [45] has been presented where NeuralTextures based learning approach has been utilized.

In the last few years, the research advancement in the development of GANs (Generative Adversarial Network) [16] is remarkable. GANs are comprised of two models, a generative model and a discriminative model, that are estimated through the adversarial process. Both models compete with each other to minimize their loss function in a fashion that can be interpreted as a minimax two-player game. GANs have been proven to be very effective for facial reenactment to generate realistic examples across a wide range of domains. The authors employ conditional GANs (cGANs) as a solution to image-to-image translation problems in their work [46] that is identical to the techniques from pix2pix software. Another work, Pix2pixHD [47], where authors proposed a successful high-resolution image generation approach using multi-scale cGANs with a perpetual loss. Wu *et al.* proposed ReenactGAN [48] to transfer both mouth and expression to the target face by mapping the source face boundary latent to the target face's boundary latent with a transformer. Eventually, reenacted target face is generated in a fake video with a target-specific decoder.

Zhang *et al.* [49] proposed a one-shot approach to generated reenacted faces using only a single source image. The authors presented an auto-encoder-based model that can learn a latent representation of both the source and target face representation. A similar but recent one-shot face reenactment model, FaR-GAN [50] has been proposed by Hao *et al.* .

2.1.5 Other Generation Methods Other than previously discussed approaches, some works with different methodologies might be classified as DeepFake generation approaches. This category includes lip-syncing, inpainting, style transfer, super-resolution, etc.

Lip-syncing DeepFake video generation approach produces a video of a target person in a fashion that the mouth and lip movement in the output video is synchronized with a source audio input. This synchronized mouth region movement makes the generated video realistic. Fan *et al.* introduced a deep bidirectional Long short-term memory (LSTM) based approach [51] for audio/visual modeling to develop a photo-real talking head system. LSTM is a subset of recurrent neural network (RNN) architecture that can model sequential data where long-term dependencies need to be considered. LSTM is widely adopted in prediction from health data [52], natural language processing (NLP) [53], next video frame prediction [54] etc. LSTM model with other DL models can learn and predict the lip movement from the input source audio file.

The image inpainting approach involves reconstructing the missing or incomplete part of images or videos. Yu *et al.* presented a well-known work, ContextAtten [55] in 2018 in image inpainting. The most common issue in the previous image inpainting is distorted structures or blurry texture in the manipulated image. Later, this issue is addressed in ContextAtten. SC-FEGAN [56] by Jo

et al. is an image editing work focusing on utilizing a relatively free-form user input in terms of color and shape.

Single image super-resolution (SISR) task might be considered as a variation of DeepFake generation. Dai *et al.* proposed the Second-order attention network (SAN) [57] for more effective feature expression and correlation learning. In this work, the authors focus on feature correlation instead of using a model with deep architecture. Karnewar *et al.* proposed MSGGAN [58] for high-resolution image synthesis with a goal of achieving well convergence on a variety of image datasets. For increasing convergence stability, The authors allow the gradients-flow from the discriminator to the generator at various scales.

One of the most common recognizable factors in DeepFake images is artifacts in their frequency domain. Some of the DeepFake research community try to eliminate traceable artifacts by modifying the generation procedure. SDGAN [59], WUCGAN [60] are examples of such works. In WUCGAN, a spectral regularization has been used to overcome the GANs' inability to produce real image spectral distribution due to the up-sampling method.

In Table 1, we provide summaries of the DeepFake generation works mentioned above.

Table 1: Summary of DeepFake Generation approaches. We list 26 recent approaches published in peer-reviewed journals and conferences in the table.

Works	**Methods**	**Elo Rating**	**Datasets**	**Multimedia**	
				Image	**Video**
[21]	Glow	1511	CIFAR-10, ImageNet, LSUN	✓	✗
[56]	SC-FEGAN	1489	CelebA-HQ	✓	✗
[18]	SAN	1487	LSUN, Imagenet, Faces, CIFAR-10, SVHN, MNIST	✓	✗
[23]	WGAN-GP	1435	LSUN, Google Billion Word, Swiss Roll	✓	✗
[61]	ContextAtten	1430	CelebA, CelebA-HQ, ImageNet, Places2, DTD	✓	✗
[42]	AttGAN	1426	CelebA, LFW	✓	✗
[20]	CocoGan	1400	CelebA, CelebA-HQ, LSUN, Matterport3D	✓	✗
[60]	WUCGAN	1400	FaceForensics++, CelebA, Faces, faces-HQ	✓	✓
[34]	CycleGAN	1400	Cityscapes, CMP Facade, UT Zappos	✓	✗
[40]	GANimation	1390	EmotioNet, RaFD	✓	✗
[38]	StarGAN	1387	CelebA, RaFD	✓	✗
[58]	MSGGAN	1385	CelebA-HQ, CIFAR-10, OF, LSUN, FFHQ	✓	✗
[62]	DCGAN	1337	HWDB1.0, LSUN, MNIST	✓	✗
[39]	HomointerpGAN	1335	RaFD, CelebA	✓	✗
[25]	PGGAN	1336	CelebA, LSUN, CIFAR10	✓	✗
[24]	BEGAN	1291	CelebA	✓	✗
[36]	RSGAN	N/A	CelebA	✓	✗
[50]	FaR-GAN	N/A	VoxCeleb1	✓	✗
[59]	SDGAN	N/A	FFHQ	✓	✗
[43]	STGAN	N/A	FFHQ	✓	✗
[48]	ReenactGAN	1400	CelebV, WFLW	✗	✓
[41]	StarGANV2	1400	CelebAHQ, AFHQ	✓	✗
[26]	StyleGAN	N/A	FFHQ	✓	✗
[27]	StyleGAN2	1416	FFHQ	✓	✗
[35]	FSGAN	1400	IJB-C, VGGFace2, Figaro, Forensics++, CelebA	✓	✓
[19]	CoGAN	N/A	MNIST, USPS, CelebA, RGBD, NYU	✓	✗

2.2 DeepFake Generation Dataset

A proper dataset plays a pivotal role in the deep learning model's performance. For DeepFake generation, two major types of datasets are used for different purposes: real dataset and synthesized dataset. In most cases, real datasets are used for DeepFake generation, whereas synthesized or fake datasets are used for DeepFake detection. Discussion on popular real datasets in the provided section below.

Yi *et al.* presented CASIA-WebFace [63] in 2014. The dataset includes 10,575 subjects and 494,414 images. CelebA [64] dataset was introduced by Liu *et al.* . This is a labeled version of CelebFaces [65]. In CelebA dataset, there are 10,000 subjects, a total of 200,000 images where each subject has twenty samples.

VGGFace [66] is a large dataset with 2.6 million images of 2,622 subjects. Microsoft Celeb (MS-Celeb-1M) [67] is another large image dataset, particularly for face recognition. This dataset contains 10 million face images of 100k identities. Cao *et al.* introduced VGGFace2 [68] in 2018, containing 3.31 million images of 9131 persons. Images are collected from the internet. This dataset has a wide variation in age, ethnicity, and pose. Flickr-Faces-HQ [26] was presented in 2019 by Karras *et al.* containing 70k high-resolution images.

3 DeepFake Detection

It is obvious that DeepFakes can be tremendous threats ranging from a person to the whole world. To avoid the threats, an effective set of DeepFake detection approaches are required. Subsequently, an increase amount of research is performed for developing various approaches to detect the authenticity of a still image or video content. In the past, a manipulated content was determined manually by analyzing artifacts and inconsistencies. In recent days, because of leveraging deep learning, complex and discriminative features are extracted to detect fake contents.

3.1 Detection Approaches

In this section, we review recent studies on various DeepFake detection works, based on their methodologies and extracted attributes.

3.1.1 Forensics-based Detection Recent forensics-based detection studies analyze pixel-level disparity. In addition, they provide explainable detection mechanism to determine authenticity. however, these works undergo robustness issues when the manipulated contents are generated by simple transformations.

Li *et al.* observed that the disparities between manipulated and real faces are revealed in the color components [69]. The authors proposed training a one-class classifier on real face data based on considering the disparities in the color components to detect the unknown GANs. However, they did not clarify the

performance of their approach against perturbation attacks such as image transformations.

In [70], Koopman *et al.* proposed to detect fake videos based on the unique noise pattern in the videos that is caused by the camera sensor. Rather than considering noise, the authors of [71] observed that DeepFakes usually contain inconsistent or unusual head poses with respect to expression. Hence, their work monitors facial landmarks to calculate the differences in head pose between manipulated and genuine video frames. They use the difference in estimated head pose data as a feature vector to train an SVM based classifier to predict original and DeepFakes. Unfortunately, the authors did not clarify the effectiveness of their works [70, 71] in detecting high quality DeepFakes.

In [72], Wang *et al.* leveraged the local motion features captured from real videos to identify the inconsistency of forged videos. They emphasized on the low-level features that are not feasible to be deployed in the wild where DeepFakes suffer known and unknown degradations.

Demir *et al.* [73] focused on synthetic eyes construction in deep fake videos. They generated features from eye and gaze data to train their model and compare it with complex state-of-the-art CNNs (VGG19, Inception, Xception, ResNet, and DenseNet). They claimed fake detection accuracy is around 6.5% higher than those of complex architectures without using eye or gaze information. However, considering only gaze data for detecting a synthesized content does not indicate improvement in generalization. Therefore, it is not clear that how the model will perform in detecting unseen adversary.

3.1.2 Deep Neural Network-(DNN) based Detection For DeepFakes detection in images and videos, neural network models with deep architecture outperform classical/hand-crafted approaches. DNN based models are capable of learning meaningful features from available data for effective forgery detection. Guera *et al.* [74] proposed a DeepFake video detection framework model combined with an RNN and CNN architecture to detect forged part from the input video. However, the limitation of the work is its incapability to handle videos for more than 2 seconds.

Nataraj *et al.* [75] presented a CCN model-based approach to calculate pixel co-occurrence matrices from the input image to detect image manipulation. Nguyen *et al.* [76] proposed a robust DeepFake video detector with multi-task CNN base architecture to identify and localize the manipulated portions from a video. An autoencoder and a decoder are used for the classification of manipulated content and sharing the extracted features for segmentation and shape reconstruction, respectively. However, the accuracy of this model declines for unseen examples that can be considered as a limitation of the work. To address this accuracy degradation-related limitation, a Forensic Transfer (FT) based CNN approach [77] for DeepFake detection was proposed by Stehouwer *et al.* . Marra *et al.* [78] presented an approach based on incremental learning for GAN-generated fake image detection. This work focuses on the classification and detection of a new type of GAN-generated images with high accuracy. This ap-

proach is more generalized and robust against detecting unseen GAN-generated examples. However, This procedure still needs some information about the new GAN architecture, which affects the practicality of the work.

Table 2: We present Summary of existing top-25 notable DeepFake detection work published in peer-reviewed journals and conferences in the table. We report only the highest achieved resultant metrics. ACC, PRE, AUC, and EER denotes accuracy, precision, area under the curve, and equal error rate, respectively.

Works	**Methods**	**Performance**	**Datasets**	**Multimedia**	
				Image	**Video**
[69]	One-class	ACC: 0.98	Self-built	✓	✗
[79]	VAE	ACC: 0.98	FF++	✗	✓
[80]	CNN	ACC: 0.96	FF, FF++, Celeb-DF	✗	✓
[81]	S-MIL	ACC: 0.83	FF++, Celeb-DF, DFDC	✗	✓
[82]	RNN	AUC: 0.99	FF++, Celeb-DF, DFDC	✗	✓
[83]	CNN	AUC: 0.99	UADFV, Celeb-DF, FF++	✗	✓
[84]	CNN	ACC: 0.98	Celeb-DF, UADFV, DFFD	✓	✗
[85]	DNN	AUC: 0.96	TIMIT, DFDC	✗	✓
[86]	CNN	PRE: 1.0	FF++	✓	✓
[87]	HRnet	ACC: 0.95	UADFV, Celeb-DF, FF++	✗	✓
[88]	CNN	ACC: 0.98	FF++	✗	✓
[76]	CNN	ACC: 0.93	FF++	✗	✓
[89]	CNN	ACC: 0.98	FF++	✗	✓
[90]	RNN	ACC: 0.99	FF++	✗	✓
[72]	CNN	PRE: 1.0	FF++	✓	✓
[74]	RNN	ACC: 0.97	self-built	✗	✓
[70]	N/A	N/A	self-built	✗	✓
[75]	CNN	ACC: 0.99	self-built	✓	✗
[73]	CNN	ACC: 0.89	FF++	✗	✓
[91]	CNN	EER 0.13	FF++, Celeb-DF	✗	✓
[77]	CNN	ACC 1.0	Self-built	✓	✗
[92]	CNN	ACC: 1.0	Self-built	✓	✗
[93]	SVM	AUC: 0.91	Self-built	✓	✗
[78]	N/A	N/A	Self-built	✓	✗
[94]	CNN	ACC: 0.99	Self-built	✓	✗

3.1.3 GAN Redesign-based Detection In lieu of considering only audio-visual artifacts, a few number of research criticize the design limitation of existing GAN-based approaches and highlight to redesign GAN by including new artifacts.

In [93], McCloskey and Albright investigated the traditional architecture of the generator function and observed that the internal values of the generator are normalized. They claimed the normalization technique limits the frequency of the saturated pixels and makes it difficult to calculate the occurrences of saturated and underexposed pixels. They suggested to use their proposed approach as a complementary to other approaches that detect visual artifacts in the manipulated contents.

Zhang *et al.* investigated how the generalization ability of the existing detectors are impacted due to the existing upsampling design related artifacts [92]. They also noted that upsampling design is generic in GAN pipelines. Hence, they proposed a new signal processing analysis and redesigned the classifier accordingly. In addition, they developed a simulator framework, AutoGAN that simulates the common generation pipeline shared by a large class of popular GAN models [92]. AutoGAN simulates the GAN generation pipeline and generates (simulated) fake images that can be used in training any classifier without the burden of accessing pre-trained GANs.

In [94], Yu *et al.* proposed GAN fingerprint artifact for classifying the images and also determining the source of a target images. Although an insignificant amount of differences yield a distinct fingerprint, the fingerprints is vulnerable (*i.e.,* tempering) to perturbation attacks such as image transformation, blur, JPEG compression, and so on.

3.1.4 Visual and Audio Inconsistency-based Detection Mittal *et al.* [85] work has addressed the essentiality of multimodal approach for DeepFake detection. They proposed an approach combining two modalities: the audio(speech) and video (face) to extract emotional features from both modalities to detect any kinds of counterfeit in the input video. This approach won't work if multiple persons are present in one video.

In [95], the authors tried to buck forgery in the realm of videos and images, by inspecting/traceback the source/mechanism of a given DeepFake image. ML tools are not always enough to combat this kind of problem. In addition, current robust DeepFake detection systems are vulnerable to adversarial images/videos. For these reasons, instead of building a robust DeepFake image/video detection system, it is more effective and scalable to find the associated generative model. With the help of a trusted third party, we can restrict/limit the malicious purpose of usage of this model. But deniability, misattribution to the original developer still a problem of the attribution approach. In easy terms, we can mention this attribution process as *Traitor Tracing*. This system will enforce accountability among model developers.

Along side with DeepFake video generation, audio spoofing is another way of character assassination of a public figure. Chintha *et al.* [91] addressed this problem by finding inconsistencies in audio and video modalities. To this end, the authors leveraged XceptionNet architecture for facial feature extraction and stacked convolutional layers to generate audio embedding features. Our analysis suggests that the combination of spoof audio and fake video detection is prone

to achieve better generalization that indicates robustness in detecting unknown adversaries.

3.1.5 Other Notable Detectors Rashmiranjan *et al.* (2021) in [96], investigated a technique involving Euler video magnification (EVM) process extracting features using three techniques (SSIM, LSTM, Heart Rate Estimation) to train models to classify counterfeit and unaltered videos. This technique uses spatial decomposition and temporal filtering on video data to highlight and magnify hidden features such as pulse of skin or subtle motions.

Fernandes *et al.* applied similar technique in [97], where they used EVM in color-based photoplethy smography (PPG) to identify blood volume fluctuations by shining light of certain wavelength onto the skin and measuring changes in light assimilation of the oxygenated blood which is in turn measures the heart rate. On the other hand, in this work, the authors applied both the EVM based color and movement amplification on videos to distinguish between original and fake videos. The results using SSIM technique when used a range of standard machine learning shows below 82% accuracy achieved by the best performing submission to the DFDC while the results using LSTM technique establishes apparent setback to the idea of using EVM for DeepFake detection. Overall, even though the color and spatial aspects of EVM were tested as possibilities for a number of classification models, the authors used accuracy as a metric though it is not known to be great metric for evaluation when imbalanced datasets are used which can be improved.

Hussain *et al.* [98] and Carlini *et al.* [99] discussed the vulnerability of current DeepFake detectors in light of both the whitebox and blackbox attacking approach. Carlini *et al.* [99] also demonstrated that a novice attacker can effectively conduct a blackbox attack without having any information regarding classifier and can reduce classifier's AUC to 0.22.

DeepTag [100] is another digital watermarking-based proactive system to combat DeepFake problems. This system finds the source of a DeepFaked image with an embedded message associated with the original image. This system works better against the dynamic image transformation and reconstruction of images by the DeepFake process. By blocking the confirmed DeepFake, this system also helps to stop spreading misinformation on the different social media platforms. According to their approach, the embedded message has to avoid the manipulated region. Even though the authors addressed this problem, they did not provide any solution on this.

DeepFake detection becomes more challenging when multiple faces are observed in a video frame. Charitidis *et al.* [101] tried to solve this problem with the improvement of preprocessing step. They pruned a cluster of facial data that carries less significance. This approach makes DeepFake detection process fast and it can be used on top of any existing DeepFake detection system. This approach still has one problem, their preprocessing approach can discard less prevalent but significant data from the datasets.

Table 2 shows the summary of the aforementioned DeepFake detection works and corresponding dataset information.

3.2 DeepFake Detection Dataset

The DeepFake Detection Challenge (DFDC) Preview Dataset
In [15], Dolhansky *et al.* introduced a preview of DFDC dataset containing 5,000 videos that featured two facial modification algorithms where the actors were in agreement to use and manipulate their likeliness. To ensure visual variability, diversity in several axes (gender, skin tone, age) and arbitrary background was considered. A reference performance baseline was provided in terms of specific metrics that was defined and tested on two existing models for detecting DeepFakes. The initial baseline consists of the performance check of three simple detection models. The first model is trained to detect low-level image and the other two models were trained on the FaceForensics++ dataset [102] and evaluated as implemented in [45]. All performances of these three sample detection models were analyzed using precision, recall, and the logarithmic scale of weighted precision to detect half, most, or nearly-all DeepFakes.

The Celeb-DF Dataset
At least until the year 2019, DeepFake datasets included low visual quality and had little to no resemblance to DeepFake videos found online. The work presented in [103] has constructed a large scale DeepFake video dataset called *Celeb-DF* that includes a total of 5,639 high-quality DeepFake videos, corresponding to more that 2 million frames from publicly available YouTube video clips of 59 celebrities of diverse genders, ages, and ethnic groups using improved synthesis process. The video quality in *Celeb-DF* with very few notable visual artifacts have significant differences with then available DeepFake videos available online that included low-quality synthesized faces, visible splicing boundaries, color mismatch, and inconsistencies in synthesized face orientation etc. The overall quality of the videos were enhanced in terms of improving low resolution of synthesized faces, color mismatch, inaccurate face masks, and temporal flickering. The authors also presented a comprehensive evaluation with 9 DeepFake detection methods and datasets considered making the most comprehensive study of DeepFake detection available by then. Overall, this *Celeb-DF* dataset has helped lowered the gap in the video quality between the actual and DeepFake datasets that can be found online with a possibility of enlarging *Celeb-DF* and further enhancing the visual quality including the running efficiency.

The FaceForensics++ dataset
Rossler *et al.* in [102] generated a large-scale dataset with an automated benchmark based on classical computer-graphics and learning-based based methods such as DeepFakes [104], FaceSwap [105], NeuralTextures [106], and Face2Face [107]. This benchmark contains a hidden test set of an order magnitude larger than comparable, publicly available dataset including 1.8 million manipulated images extracted from 1,000 real videos and target ground truth to enable supervised learning. The authors conducted a thorough study of data-driven forgery detectors and showed that the use of domain specific information in conjunction

with a XceptionNet classifier improves the detection with an unprecedented accuracy. This work also presented ways to automatically detect any forms of facial identity and facial expression manipulations with an automated benchmark consisting with random compression and dimensions.

The UADFV dataset

Up until the revelation of the popular software "DeepFake"that used generative adversary networks (GANs), since any form of manipulation of videos/images involved huge time consumption for editing operations, realistic high quality fake videos were not widespread. Due to this software, ample of high-quality fake video flooded the internet and thus detecting such videos became important. In [108], the authors used 50 YouTube videos that lasted 30 seconds each representing one individual with at least one blinking occurred, to form the Eye Blinking Video (EBV) dataset. The left and right states of each frame were annotated with a user-friendly annotation tool. The training dataset that was used in this work is CEW [109] which includes 1,793 images of closed eyes and 1,232 images of open eyes to train front-end CNN model, 40 videos as training set for the overall Long-term Recurrent Convolutional Networks (LRCN) model [110] and 10 videos as the testing set. This LRCN method was shown to exhibit best performance 0.99 compared to other methods such as Eye Aspect Ratio (EAR) [111] with performance 0.79 and CNN with 0.98.

4 Challenges and Opportunities for Future Research Direction

We review over 60 articles published either in peer-reviewed journals and conferences or posted on arXiv regarding DeepFake generation and detection. In this section, we describe our observations including challenges, limitations, and new research scopes, after reviewing the studied articles. This will contribute on the future research in creating more realistic and detection-evasive DeepFake, and also sophisticated DeepFake detection model.

4.1 Findings in DeepFake Generation

– We notice low resolution and poor quality in the output image irrespective to any existing DeepFake generation work. Currently, generation of high-resolution and sharp images is difficult since such image makes the job easier to differentiate it from training images [112]. A deeper analysis claims that this causes spike in the gradient problem and affects training stability [113]. To mitigate the challenge, PGGAN is proposed to grow generator and discriminator progressively. It starts from low-resolution images and gradually, adds new layers for higher-resolution (*i.e.,* details) as the training progresses [113]. However, PGGAN is still in premature stage and capable of generating only (1024×1024) size images.

- The attribute manipulation methods are limited as these can only change the properties followed by the training set. Therefore, such an attribute manipulation method is needed that could capture attributes are independent attributes to the training set.
- In most of the cases, identity swap and expression swap do not consider the continuity of the videos. They neither consider gesture nor physiological signals such as eye blink, breathing frequency, heart beat, and so on.
- The fake datasets are expanded only considering the diversity of the content-related factors such as age, gender, background, and so on. Our investigation conform factors such as added noise (quality degradation), Gaussian blue, JPEG compression, contrast change, and so on can enhance the diversity in the dataset. DeeperForensics-1.0 [114] dataset offers image-level degradations but it is added artificially by post-processing. We expect natural image/video-level degradations (*i.e.,* over/underexposed photos, bit-rate variations, choices of codec) in the future generation of the dataset.

4.2 Findings in DeepFake Detection

- Most of the existing works generate image dataset to evaluate the effectiveness of their approaches leveraging various GANs. A large portion of the works do not unveil the details about the datasets that used in evaluation. Hence, the quality of the generated forged images still remains unknown. On the contrary, these works claim their competitive results in detecting various synthesized images built on their own. We emphasize on the development of public GAN-synthesized fake image dataset.
- For the sake of performance evaluation, existing works implement simple baselines (*e.g.,* vanilla DNN-based methods) and compare it with their works rather than considering state-of-the-art ones. Claiming the superiority of their works by comparing with naïve approaches indicate biased evaluation. We expect that future works should be comparable to state-of-the-art works so that we can understand effectiveness of the proposed work.
- The aim of DeepFake detection research is to develop more robust and generalized approaches. Subsequently, the research community is trying their level best to come up with effective approaches. However, the recent works are simply evaluated on simple DeepFake video datasets, such as FaceForensics++ (FF++). We emphasize that future works should focus more on challenging datasets for acceptable performance evaluation.
- Almost all of the existing studies report their experimental results by merely considering the detection accuracy without reporting other popular metrics such as precision, recall, and the relation with the quality of DeepFakes. To conduct an acceptable performance evaluation, a comprehensive experimental result set is mandatory. A robust set of experiments should contain the result of various effective metrics. Currently, there does not exist any metric that can measure DeepFake quality. We hope, in future, the researchers would come up with a new metric for measuring the quality of DeepFakes.

– Detecting the emerging unknown DeepFakes is crucial in today's world. Hence, developing a practical DeepFake detector that is deployable in the wild is a necessity. Towards developing an effective DeepFake detector, we observe a set of key factors that are: (1) advance generalization capabilities, robust against various attacks(*e.g.,* iamge/video transformations, adversarial attacks), and presenting explainable DeepFake detection result. Unfortunately, in reviewing the recent DeepFake detection articles, we find that the researchers simply ignore to evaluate the capabilities of their works from the aforementioned perspectives.

5 Conclusion

Due to great progress on generative deep learning algorithm in recent years, nowadays it has become a real challenge to identify the authenticity of any visual content found online [115]. The aim of creating the synthesized contents is either for malicious intent or just for fun. To resist any unexpected scenarios such as creating a manipulated content of important persons (*e.g.,* political leaders, celebrities) or generate synthesized contents for a useful purpose, the current DeepFake research community needs to consider the existing published articles both in DeepFake generation and detection to plan for extensive research efforts in the future. In light on this and to make our understanding better, in this current work, our investigation shows that in recent years deep learning research community have been trying to solve two large research domains including DeepFake detection and generation related to DeepFake image and video contents. We have shed light on these domains by discussing state-of-the-art research works. We also try to depict how the research community shifts their attention from feature-based DeepFake detection to feature agnostic and policy-based approaches to combat evasion of DeepFake detection. We also provide comprehensive descriptions of different prominent datasets to facilitate researchers to determine their next research direction.

Bibliography

[1] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets.

[2] Deepfakes porn has serious consequences. https://www.bbc.com/news/technology-42912529, Accessed April 1, 2021.

[3] Deepfake Porn Nearly Ruined My Life. https://www.elle.com/uk/life-and-culture/a30748079/deepfake-porn, Accessed April 1, 2021.

[4] 12 deepfake examples that terrified and amused the internet. https://www.creativebloq.com/features/deepfake-examples, Accessed April 1, 2021.

[5] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. Deepfake: improving fake news detection using tensor decomposition-based deep neural network. *The Journal of Supercomputing*, pages 1–23, 2020.

[6] A guy made a deepfake app to turn photos of women into nudes. It didn't go well. https://www.vox.com/2019/6/27/18761639/ai-deepfake-deepnude-app-nude-women-porn, Accessed April 1, 2021.

[7] Chinese deepfake app Zao sparks privacy row after going viral. https://www.theguardian.com/technology/2019/sep/02/chinese-face-swap-app-zao-triggers-privacy-fears-viral, Accessed April 1, 2021.

[8] Siwei Lyu. Deepfake detection: Current challenges and next steps. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6, 2020.

[9] Luca Guarnera, Oliver Giudice, Cristina Nastasi, and Sebastiano Battiato. Preliminary forensics analysis of deepfake images. In *Proceedings of International Annual Conference (AEIT)*, pages 1–6, 2020.

[10] Felix Juefei-Xu, Run Wang, Yihao Huang, Qing Guo, Lei Ma, and Yang Liu. Countering malicious deepfakes: Survey, battleground, and horizon. *arXiv preprint arXiv:2103.00218*, 2021.

[11] Mousa Tayseer Jafar, Mohammad Ababneh, Mohammad Al-Zoube, and Ammar Elhassan. Forensics and analysis of deepfake videos. In *Proceedings of the 11th International Conference on Information and Communication Systems (ICICS)*, pages 053–058, 2020.

[12] Loc Trinh, Michael Tsang, Sirisha Rambhatla, and Yan Liu. Interpretable deepfake detection via dynamic prototypes. *arXiv preprint arXiv:2006.15473*, 2020.

[13] S.3805 - Malicious Deep Fake Prohibition Act of 2018 . https://www.congress.gov/bill/115th-congress/senate-bill/3805, Accessed April 1, 2021.

[14] Help us shape our approach to synthetic and manipulated media. https://blog.twitter.com/en_us/topics/company/2019/synthetic_manipulated_media_policy_feedback.html, Accessed April 1, 2021.

[15] Brian Dolhansky, Russ Howes, Ben Pflaum, Nicole Baram, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

[16] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[18] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[19] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. *arXiv preprint arXiv:1606.07536*, 2016.

[20] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. Coco-gan: Generation by parts via conditional coordinating. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4512–4521, 2019.

[21] Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018.

[22] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of international conference on machine learning*, pages 214–223. PMLR, 2017.

[23] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. *arXiv preprint arXiv:1704.00028*, 2017.

[24] David Berthelot, Thomas Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[25] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[28] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[30] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[31] Samet Akcay, Mikolaj E Kundegorski, Chris G Willcocks, and Toby P Breckon. Using deep convolutional neural network architectures for object classification and detection within x-ray baggage security imagery. *IEEE transactions on information forensics and security*, 13(9):2203–2215, 2018.

[32] Md Imran Hossen, Yazhou Tu, Md Fazle Rabby, Md Nazmul Islam, Hui Cao, and Xiali Hei. An object detection based solver for google's image recaptcha v2. In *23rd International Symposium on Research in Attacks, Intrusions and Defenses ({RAID} 2020)*, pages 269–284, 2020.

[33] FaceSwap. https://github.com/deepfakes/faceswap. 2016.

[34] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[35] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7184–7193, 2019.

[36] Ryota Natsume, Tatsuya Yatagawa, and Shigeo Morishima. Rsgan: face swapping and editing using face and hair representation in latent spaces. *arXiv preprint arXiv:1804.03447*, 2018.

[37] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019.

[38] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[39] Ying-Cong Chen, Xiaogang Xu, Zhuotao Tian, and Jiaya Jia. Homomorphic latent space interpolation for unpaired image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2408–2416, 2019.

[40] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 818–833, 2018.

[41] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[42] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019.

[43] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3673–3682, 2019.

[44] Justus Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: real-time face capture and reenactment of rgb videos. *ArXiv*, abs/2007.14808, 2019.

[45] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision (ICCV)*, 2019.

[46] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.

[47] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans, 2018.

[48] W. Wu, Yunxuan Zhang, Cheng Li, Chen Qian, and Chen Change Loy. Reenactgan: Learning to reenact faces via boundary transfer. *ArXiv*, abs/1807.11079, 2018.

[49] Yunxuan Zhang, S. Zhang, Yue He, Cheng Li, Chen Change Loy, and Z. Liu. One-shot face reenactment. *ArXiv*, abs/1908.03251, 2019.

[50] Hanxiang Hao, Sriram Baireddy, A. Reibman, and E. Delp. Far-gan for one-shot face reenactment. *ArXiv*, abs/2005.06402, 2020.

[51] Bo Fan, Lijuan Wang, Frank K Soong, and Lei Xie. Photo-real talking head with deep bidirectional lstm. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4884–4888. IEEE, 2015.

[52] Md Fazle Rabby, Yazhou Tu, Md Imran Hossen, Insup Lee, Anthony S Maida, and Xiali Hei. Stacked lstm based deep recurrent neural network with kalman smoothing for blood glucose prediction. *BMC Medical Informatics and Decision Making*, 21(1):1–15, 2021.

[53] Sm Zobaed, Md Enamul Haque, Md Fazle Rabby, and Mohsen Amini Salehi. Senspick: sense picking for word sense disambiguation. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 318–324. IEEE, 2021.

[54] Matin Hosseini, Anthony S Maida, Majid Hosseini, and Gottumukkala Raju. Inception lstm for next-frame video prediction (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13809–13810, 2020.

[55] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[56] Youngjoo Jo and Jongyoul Park. Sc-fegan: face editing generative adversarial network with user's sketch and color. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1745–1753, 2019.

[57] Tao Dai, Jianrui Cai, Yongbing Zhang, Shu-Tao Xia, and Lei Zhang. Second-order attention network for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11065–11074, 2019.

[58] Animesh Karnewar and Oliver Wang. Msg-gan: Multi-scale gradients for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7799–7808, 2020.

[59] Steffen Jung and Margret Keuper. Spectral distribution aware image generation. *arXiv preprint arXiv:2012.03110*, 2020.

[60] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7890–7899, 2020.

[61] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[62] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, and Zhen Wang. Multi-class generative adversarial networks with the l2 loss function. *arXiv preprint arXiv:1611.04076*, 5:00102, 2016.

[63] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[64] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[65] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Hybrid deep learning for face verification. In *Proceedings of the IEEE international conference on computer vision*, pages 1489–1496, 2013.

[66] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. 2015.

[67] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016.

[68] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

[69] Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang. Identification of deep network generated images using disparities in color components. *Journal of Signal Processing*, 174:107616, 2020.

[70] Marissa Koopman, Andrea Macarulla Rodriguez, and Zeno Geradts. Detection of deepfake video manipulation. In *Proceedings of the 20th Irish machine vision and image processing conference (IMVIP)*, pages 133–136, 2018.

[71] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *Proceedings of international Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.

[72] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8695–8704, 2020.

[73] Ilke Demir and Umur A Ciftci. Where do deep fakes look? synthetic face detection via gaze tracking. *arXiv preprint arXiv:2101.01165*, 2021.

[74] David Güera and Edward J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[75] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, BS Manjunath, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, and Amit K Roy-Chowdhury. Detecting gan generated fake images using co-occurrence matrices. *Electronic Imaging*, 2019(5):532–1, 2019.

[76] Huy H Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8. IEEE, 2019.

[77] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018.

[78] Francesco Marra, Cristiano Saltori, Giulia Boato, and Luisa Verdoliva. Incremental learning for the detection and classification of gan-generated images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2019.

[79] Hasam Khalid and Simon S Woo. Oc-fakedect: Classifying deepfakes using one-class variational autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 656–657, 2020.

[80] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[81] Xiaodan Li, Yining Lang, Yuefeng Chen, Xiaofeng Mao, Yuan He, Shuhui Wang, Hui Xue, and Quan Lu. Sharp multiple instance learning for deepfake video detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1864–1872, 2020.

[82] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *European Conference on Computer Vision*, pages 667–684. Springer, 2020.

[83] Disheng Feng, Xuequan Lu, and Xufeng Lin. Deep detection for face manipulation. In *International Conference on Neural Information Processing*, pages 316–323. Springer, 2020.

[84] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5781–5790, 2020.

[85] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: A deepfake detection method using audio-visual affective cues. *arXiv preprint arXiv:2003.06711*, 2020.

[86] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *European Conference on Computer Vision*, pages 103–120. Springer, 2020.

[87] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5001–5010, 2020.

[88] Michail Tarasiou and Stefanos Zafeiriou. Extracting deep local features to detect manipulated images of human faces. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1821–1825. IEEE, 2020.

[89] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018.

[90] Ekraam Sabir, Jiaxin Cheng, Ayush Jaiswal, Wael AbdAlmageed, Iacopo Masi, and Prem Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, 3(1), 2019.

[91] Akash Chintha, Bao Thai, Saniat Javid Sohrawardi, Kartavya Bhatt, Andrea Hickerson, Matthew Wright, and Raymond Ptucha. Recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1024–1037, 2020.

[92] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2019.

[93] Scott McCloskey and Michael Albright. Detecting gan-generated imagery using saturation cues. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4584–4588. IEEE, 2019.

[94] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7556–7566, 2019.

[95] Baiwu Zhang, Jin Peng Zhou, Ilia Shumailov, and Nicolas Papernot. On attribution of deepfakes, 2021.

[96] Rashmiranjan Das, Gaurav Negi, and Alan F. Smeaton. Detecting deepfake videos using euler video magnification. *arXiv preprint arXiv:2101.11563*, 2020.

[97] Steven Fernandes, Sunny Raj, Eddy Ortiz, Iustina Vintila, Margaret Salter, Gordana Urosevic, and Sumit Jha. Predicting heart rate variations of deepfake videos using neural ode. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[98] Shehzeen Hussain, Paarth Neekhara, Malhar Jere, Farinaz Koushanfar, and Julian McAuley. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3348–3357, 2021.

[99] Nicholas Carlini and Hany Farid. Evading deepfake-image detectors with white-and black-box attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 658–659, 2020.

[100] Run Wang, Felix Juefei-Xu, Qing Guo, Yihao Huang, Lei Ma, Yang Liu, and Lina Wang. Deeptag: Robust image tagging for deepfake provenance. *arXiv preprint arXiv:2009.09869*, 2020.

[101] Polychronis Charitidis Giorgos Kordopatis-Zilos Symeon and Papadopoulos Ioannis Kompatsiaris. Investigating the impact of pre-processing and prediction aggregation on the deepfake detection task.

[102] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019.

[103] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020.

[104] Deepfakes github. https://github.com/deepfakes/faceswap. Accessed: 04/11/2021.

[105] FaceSwap github. https://github.com/MarekKowalski/FaceSwap/. Accessed: 04/11/2021.

[106] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[107] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[108] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. *arXiv preprint arXiv:1806.02877*, 2018.

[109] Fengyi Song, Xiaoyang Tan, Xue Liu, and Songcan Chen. Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients. *Pattern Recognition*, 47(9):2825–2838, 2014.

[110] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[111] Jan Cech and Tereza Soukupova. Real-time eye blink detection using facial landmarks. *Cent. Mach. Perception, Dep. Cybern. Fac. Electr. Eng. Czech Tech. Univ. Prague*, pages 1–8, 2016.

[112] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*, pages 2642–2651. PMLR, 2017.

[113] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *Proceedings of international Conference on Learning Representations*, 2018.

[114] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2889–2898, 2020.

[115] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41, 2021.