Challenges and Applications of Face Deepfake

Lamyanba Laishram, Md. Maklachur Rahman, and Soon Ki Jung(✉)

School of Computer Science and Engineering, Kyungpook National University,
Daegu, Republic of Korea
{yanbalaishram,maklachur,skjung}@knu.ac.kr

Abstract. With the development of Generative deep learning algorithms in the last decade, it has become increasingly difficult to differentiate between what is real and what is fake. With the easily available "Deepfake" applications, even a person with less computing knowledge can also produce realistic Deepfake data. These fake data have many benefits while on the other hand, it can also be used for unethical and malicious purposes. Deepfake can be anything fake data generated by using deep learning methods. In this study, we focus on Deepfake with respect to face manipulation. We represent the currently used algorithms and datasets are represented for creating Deepfake. We also study the challenges and the real-world applications in which the benefits, as well as the drawbacks of using Deepfake, are being pointed out.

Keywords: Deepfake · DeepFake creation · Faceswap · Face attribute editing · Deepfake dataset

1 Introduction and Motivation

Face manipulation and editing have been used in film industry for many years. The editing of facial expression of the actor is done with the changes in the dialogues of the film actors, which are sometimes hard for the actors to perform and also swapping faces between the stunt performer and the real actors of the film. However, the face manipulation is not limited to the film industry, it is also used in many other areas.

With the introduction of Generative Adversarial Network (GAN) [34] in 2014, the hype or fake vs original data is on the top. Deepfake is a powerful computer vision technique which is used to manipulate or generate fake content. The term "Deepfake" obtained from the words "deep learning" and "fake" uses the techniques of machine learning and artificial intelligence to create new and fake contents but realistic looking and difficult to distinguished whether it is fake or real. 'Deepfake with face manipulation' is a part of Deepfake where it mainly focus on the manipulation and editing of the facial part of an individual. The method superimpose faces of a target on the faces of the source and create an image or a video of the target person doing things that they actually do not do.

H. Jeong and K. Sumi (Eds.): IW-FCV 2021, CCIS 1405, pp. 131–156, 2021.
https://doi.org/10.1007/978-3-030-81638-4_11

The common benefits of this technology are their creative and productive applications in areas of education, marketing, art and multimedia [3,8,20,22,23]. With all the positive applications of this technology, the potential of malicious use is of a great concern, as the technology becomes more refine over time. The term 'Deepfake' is first used by an anonymous user of Redditor in late 2017 by posting videos of celebrity in porn videos, which is a combination of a celebrities face and body of a porn actor fused together [25]. The vulnerable victims of these malicious activities are the celebrities and political figures as their face data can be easily collected for creating of the Deepfake videos.

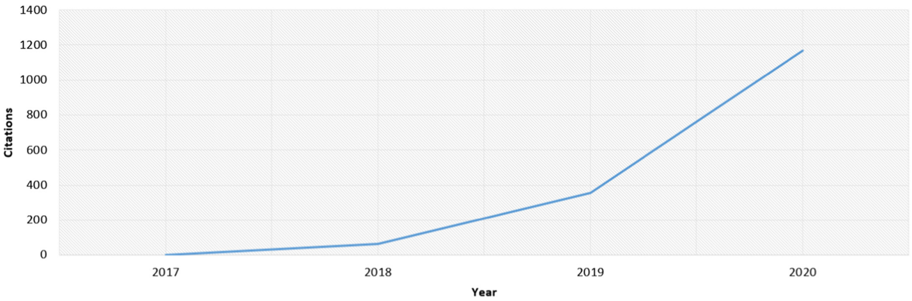

Fig. 1. Number of cited papers related to keyword "Deepfake" from [12]. The graph represents the number of papers from 2017 till the end of 2020.

Since 2017, the Deepfake technique became popular and skyrocketed in the academic community too. The technology has been advancing very rapidly and the Deepfake videos are becoming very realistic and difficult to be identified as fake videos. From the data obtained [12], the number of research papers increased from 3 to over 1000 during 2017 to 2020 as shown in Fig. 1. With the creation of realistic Deepfake contents, there is a need for the detection also. The growing interest of fake face detection is observed through the increasing number of workshops in top conferences [7,27] and in recent competition NIST's MFC2018 [21] and Facebook's DFDC [13]. At the same time the publicly available tools are also on the rise which is represented in the Table 1.

To understand the trending threats from Deepfake and in order to reduce them, we need to have a brief overview of what deepfake is. In this paper, we focus on the Deepfake with respect to the human face manipulation and edition. The main goals are (1) to identify current Deepfake creation techniques, (2) the types of database used in Deepfake creation (3) recent Deepfake approaches and the challenges, (3) discussing the future area of Deepfake.

2 Database Collection and Analysis

In this section we first reviewed the publicly available data collection in the Deepfake community used in training and testing of Deepfake creation and detection

Table 1. List of publicly available Deepfake tools.

Tools	Link	Type
DeepFaceLab	https://github.com/iperov/DeepFaceLab/	Open source
faceswap	https://github.com/deepfakes/faceswap/	Open source
FaceSwap	https://github.com/MarekKowalski/FaceSwap/	Open source
faceswap-GAN	https://github.com/shaoanlu/faceswap-GAN/	Open source
Fake	https://www.fakeapp.com/	App
FakeApp	https://fakeapp.softonic.com/	App
DeepFake_tf	https://github.com/StromWine/DeepFake_tf	Open source
Faceswap web	https://faceswapweb.com/	Website
dfaker	https://github.com/dfaker/df	Open source
fewshot-face-translation-GAN	https://github.com/shaoanlu/fewshot-face-translation-GAN	Open source
Reflect	https://reflect.tech/	Website
FaceSwap online	https://faceswaponline.com/	Website
Face Swap Live	http://faceswaplive.com/	Website
FakeApp 2.2.0.	https://www.malavida.com/en/soft/fakeapp/	App

methods and followed by the discussion on the characteristics of each set with the advantages and disadvantages thereof.

2.1 Data Collection

With the rise in the creation of manipulated faces in recent years, the key aspect of developing a good Deepfake system is the data collection used in training. However, Data collections are mostly overlooked, but they are extremely important. Based on the type of method, Data collection can be divided into three sections: only real database, real and fake database and only fake database. Table 2 shows all the Deepfake face database and their respective links.

Only Real Face Database. This database is made by the collection of only real faces found in internet. This data collection is used either in training or testing of different methods with respect to Deepfake. Currently, the main training set available are CelebFaces Attributes Dataset (celebA) [56], Radboud face database (RaFD) [49], Flickr-Faces-HQ (FFHQ) [17], CelebA-HQ [41] and CelebAMask-HQ [50].

celebA is a large-scale face attributes dataset with more than200K celebrity images, each with 40 attribute annotations. The images in this dataset cover large pose variations and background clutter. CelebA has large diversities, large quantities, and rich annotations. This database contains 10177 number of identities with 5 landmark location and 40 binary attributes annotations for each images.

RaFD is a high quality face database which contains an image set of 8 emotional expression. This 8 expressions were collected from 67 individuals including Caucasian males and females, Caucasian children, both boys and girls, and

Moroccan Dutch males. Anger, disgust, fear, happiness, sadness, surprise, contempt, and neutral are the expressions. This database also provides 3 gaze directions: looking left, front and right.

FFHQ dataset consists of 70,000 high-quality PNG images at 1024 × 1024 resolution and contains considerable variation in terms of age, ethnicity and image background. It also has good coverage of accessories such as eyeglasses, sunglasses, hats, etc.

CelebA-HQ is a high quality facial image dataset that consists of 30000 images picked from CelebA dataset. These images are processed with quality improvement to the size of 1024 × 1024.

CelebAMask-HQ is a large scale face semantic label dataset consisting of 30,000 high resolution face images from CelebA. The size of the images are 512 × 512 and 19 classes of all facial components such as skin, nose, eyes, eyebrows, ears, mouth, lip, hair, hat, eyeglass, earring, necklace, neck, and cloth.

Real and Fake Face Database. This data collections consist of both pristine and fake images. There are fake images with the corresponding real images in this collection. Currently, the main database available are UADFV [52], DeepfakeTIMIT (DF TIMIT) [46], FaceForensics ++ [69], Deepfake detection (DFD) [31], Celeb-DF [53], Deepfake detection challenge (DFDC) [31], Deeper forensics 1.0 [40], Wild deepfake [85]. Table 3 show the camparison between the different dataset for real and deefake dataset.

UADFV database consist of 49 real YouTube and 49 Deepfake videos. All the swapped faces are created with the face of actor Nicolas Cage by using FakeAPP mobile application [16]. The data provided are of resolution 294 × 500 pixels with an average of 11.14 s.

DeepfakeTIMIT is generated from the original VidTIMIT database [26]. DeepfakeTIMIT database [10] consists of 620 fake videos generated using open source GAN based approach Faceswap application [14]. The database is provided with two different qualities (i) a lower quality (LQ) with 64 × 64 size (ii) higher quality (HQ) with 128 × 128 size. Each version is generated using 10 videos from the vidTIMIT with 32 subjects which provides 320 videos for corresponding LQ and HQ.

Faceforensics++ database was introduced in 2019. It is the extension of the original Faceforensics [68] which focuses on swapping facial expressions. This database contains 1000 pristine videos downloaded from the internet (YouTube) and 1000 fake videos. The manipulated dataset is generated using two computer graphics-based approaches: Face2Face [75] and FaceSwap [15] and two learning based approaches DeepFakes [14] and NeuralTextures [74].

Deepfake detection is Google and Jigsaw DeepFake detection dataset which has 3068 DeepFake videos generated based on 363 original videos of 28 consented individuals of various genders, ages and ethnic groups. The details of the synthesis algorithm are not disclosed, but it is likely to be an implementation of the basic DeepFake maker algorithm.

Table 2. The list of all the Deepfake face database with their respective links.

Type	Database	Link	Size/type
Real database	celebA	http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html	202k/Images
	RaFD	http://www.socsci.ru.nl:8180/RaFD2/RaFD	201/Images
	FFHQ	https://github.com/NVlabs/ffhq-dataset	70k /Images
	CelebA-HQ	https://github.com/NVlabs/stylegan	30k/Images
	CelebAMask-HQ	https://github.com/switchablenorms/CelebAMask-HQ	30k /Images
Both real and fake database	UADFV	https://github.com/yuezunli/WIFS2018_In_Ictu_Oculi	90/Videos
	DeepfakeTIMIT	https://www.idiap.ch/dataset/deepfaketimit	620/Videos
	Faceforensics++	https://github.com/ondyari/FaceForensics	5k/Videos
	DFD	https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html	3.3k/Videos
	Celeb-DF	https://github.com/yuezunli/celeb-deepfakeforensics	6.2k/videos
	DFDC	https://ai.facebook.com/datasets/dfdc/	5.2k/Videos
	Deeper Forensics-1.0	https://github.com/EndlessSora/DeeperForensics-1.0	60k/videos
	WildDeepfake	https://github.com/deepfakeinthewild/deepfake-in-the-wild	7.2k/Videos
Only fake database	100k generated images	https://drive.google.com/drive/folders/100DJ0QXyG89HZzB4w2Cbyf4xjNK54cQ1	100k/Images
	100k faces	https://generated.photos/	100k/Images
	DFFD	http://cvlab.cse.msu.edu/dffd-dataset.html	300k/Images
	FSRemovalDB	https://github.com/socialabubi/iFakeFaceDB	150k/Images

Celeb-DF is created with the goal to generate a better visual quality as compared to their previous work UADFV database. This database consists of 408 real youtube videos corresponding to interviews of 59 celebrities with a diverse distribution in terms of gender, age, and ethnic group. In addition, these videos exhibit a large range of variations in aspects such as the face sizes (in pixels), orientations, lighting conditions, and backgrounds. Regarding fake videos, a total of 795 videos were created using DeepFake technology, swapping faces for each pair of the 59 subjects. The final videos are in MPEG4.0 format.

DFDC dataset is by far the largest currently and publicly-available face swap video dataset, with over 100,000 total clips sourced from 3,426 paid actors, produced with several Deepfake, GAN-based, and non-learned methods. This database was released by Facebook in collaboration with other companies and academic institutions such as Microsoft, Amazon, and the MIT for the Kaggle contest. The DFDC database considers different acquisition scenarios (i.e., indoors and outdoors), light conditions (i.e., day, night, etc.), distances from the person to the camera, and pose variations, among others.

Deeper Forensics-1.0 dataset represents a large face forgery detection dataset, with 60,000 videos constituted by a total of 17.6 million frames. The database consists of 50000 real videos and 10000 fake videos. The source videos were recorded with 100 paid actors with different genders, ages, skin colors, and nationalities. Data are recorded in a controlled indoor environment. Seven different distortions were also provided in the database.

WildDeepfake contains both real and fake videos which are collected from the internet. The video contents are diverse: variety of activities, scenes, lighting condition, compression rates, backgrounds, formats and resolution. It is a collection of 7,314 face sequences from 707 videos with annotation.

Only Fake Face Database. This data collection consists of only synthetic images, not real images. These images are generated using deep learning methods, mixing different attributes from different people to generate a new face. Currently, the main database available are 100k generated images [2], 100k faces [1], Diverse Fake Face Dataset (DFFD) [72], iFakeFaceDB [61].

100k generated images is a set of 100,000 synthetic face images. This database was generated using StyleGAN architecture [42], which was trained using the FFHQ dataset. StyleGAN [42] is an improved version of the previous popular approach ProGAN [41], which introduced a new training methodology based on improving both generator and discriminator progressively. StyleGAN proposes an alternative generator architecture that leads to an automatically different learning style corresponding to different spacial resolution.

100k faces is another face synthetic public database. This database contains 100,000 synthetic images generated using StyleGAN as well. In this database, contrary to the 100K-Generated- Images database, the StyleGAN network was trained using around 29,000 photos from 69 different models, considering face images from a more controlled scenario (e.g., with a flat background). Thus, no strange artifacts created by the StyleGAN are included in the background of the images.

DFFD is a new database comprised of publicly available datasets and images that are synthesized/manipulated using publicly available methods. Regarding the entire face synthesis manipulation, the authors created 100,000 and 200,000 fake images through the pre-trained ProGAN and Style-GAN models, respectively.

iFakeFaceDB which is the Face Synthetic Removal database (FSRemovalDB). This database comprises of a total 150,000 synthetic face images originally created through StyleGAN. Contrary to the other databases, in this database the GAN "fingerprints" produced by the StyleGAN were removed from the original synthetic fake images through the use of autoencoders, while keeping the visual quality of the resulting images. Therefore, this database presents a higher level of manipulation for the detection systems.

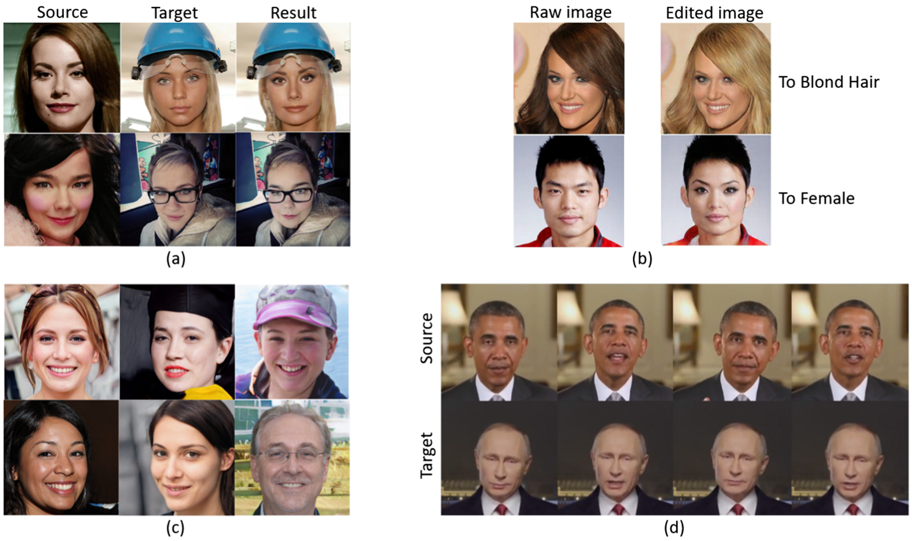

Fig. 2. Different DeepFake creation techniques; (a) face swap (b) attribute editing (c) face synthesis (d) reenactment.

2.2 Analysis

Deepfake database can be categorized into different generations as the methods for producing fake videos are continuously improved with time. More and more realistic dataset are produced and made available for the community. Some of the factors based on which the generations can be categorized are size of the database, variation and diversity of the data, pose and illumination, and quality of the image.

First Generation: Datasets which contain less then 1000 videos and the quality of the Deepfake videos are usually of low quality. The database included here are UADFV [52], DeepfakeTIMIT (DF TIMIT) [46], FaceForensics ++ [69].

These datasets contain videos from internet as source and swap face faces between the individuals. Additionally, there is no underlying consent or agreement with the individuals in the dataset.

Second Generation: The quality of the deepfake videos is much better as compared to the first generation. This category includes Deepfake detection (DFD) [31], Celeb-DF [53],WildDeepfake [85]. The consent of the individual in the database was publicly raised [71]. Some of the dataset include paid actors but are not enough for proper detection.

Third Generation: These dataset are very realistic and more diverse as compared to the previous generations. Deepfake detection challenge (DFDC) [31] and Deeper forensics 1.0 [40] are in this generation. DFDC dataset contains individuals in real world lighting conditions while Deeper forensics 1.0 contains videos taken in a control environment. Deeper forensics 1.0 provides different poses, expression and perturbations but has only 1000 real videos. DFDC has a lot of Deepfake videos produced by target/source swap.

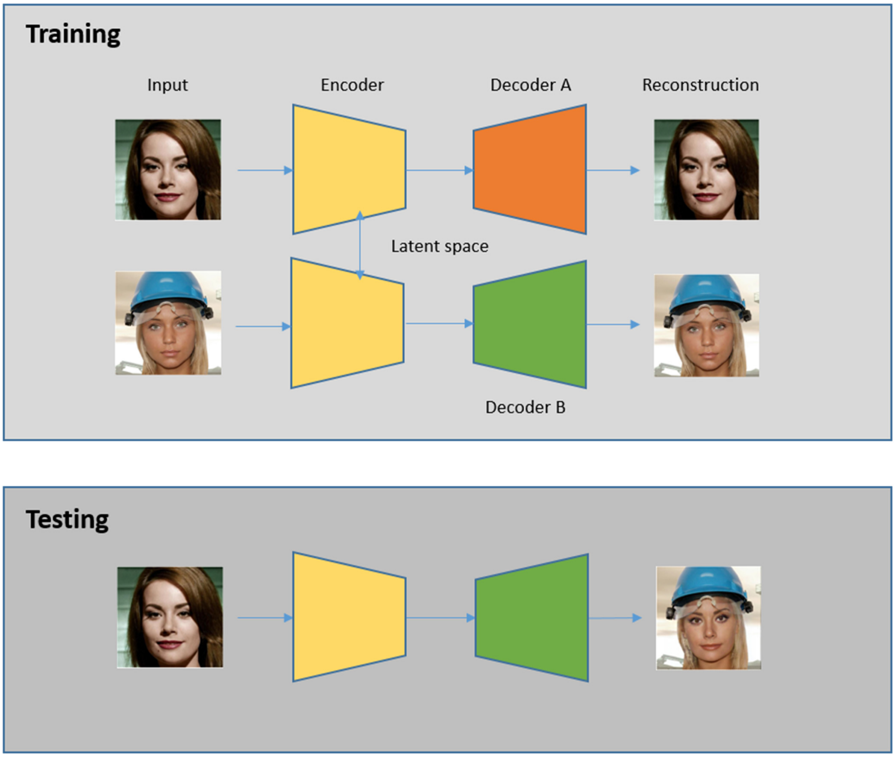

Fig. 3. Basic face swap technique using two encoder-decoder pair. Top presents the training process while other represents testing.

3 Deepfake

Deepfake is generally defined as the manipulated media which are difficult to distinguish between being pristine and being fake. The Deepfake domain is vast but in this article, we focus on the fake human faces. In the context of visual and technique, four categories are made: face replacement, facial attribute editing, face reenactment and face synthesis. Figure 2 shows the creation techniques. Considering that we have a source and a target person, the categories are discussed in the following sections.

3.1 Face Replacement

Face replacement is commonly known as "face swap". Face swap is the process of swapping a face of a person by another person's face and was first created by a Reddit user [14]. Let's consider two person, A and B. For the basic face swap creation, two autoencoder-decoder pair structure is required. The latent features from the face image are extracted using an autoencoder and then a decoder is used to reconstruct the original input face image from the latent features. One pair is first trained using a set of images for person A and another pair using the set of images for person B but the encoder's parameters are shared between the two network pairs. Both pairs have the same encoder and two different decoders. The target identity replaces the source identify while preserving headpose, exact facial expression and lighting condition. This approach replaces eyes, eyebrow, nose, mouth and sometime the contour of the face. Figure 3 show the basic face swap method.

Trending Approaches

Preprocessing: The first step is to get a collection of the face dataset. The dataset can be categoriesd into two types of input images, the source and the target images. The source image provides identity or content and a target image provides attributes, e.g., pose, expression, scene lighting and background. The goal is to render the semantic content of one face image to the style of another image. First, we need to locate the face area in an image or each frame if it is a video using a face detector. Then, the facial landmarks which carries important structural information such as eyes, noses, mouth and contours of the face area are to be located. The change of the head movement and change in the face orientation should be addressed as these changes produce artifacts in the generated face swapping. We must align the face region in a proper coordinate space by face alignment algorithm based on the landmarks obtained. The algorithm MTCNN [81] is a face detector commonly used to identify the face regions in each image or video frame. Facial keypoint or landmarks are extracted using dlib [44] but it is sometimes difficult for partial occluded face region.

Another face alignment technique is to fit a 3D shape on to the target face and modify the 3D faces to account for facial expression represented in [76].

Some of the popular 3D shapes are Basel Face Model (BFM) [63] to represent faces and the 3DDFA Morphable Model [84] for expression. Face segmentation can also be done by fitting 3D shape and it can also address the problem of partial occlusion on the face region [62]. The drawback of 3DMM swap method is that manual alignment is required for accurate fitting.

Different segmentations process are required for different methods. Some methods require face segmentation for processing [47], some use hair, face and background segmentation [60] and some use mask of the face [59]. The creation of a smooth and precise face mask based on the landmarks on eyebrow and points on cheeks and between lower lip and chin is critical in the final blending [53]. Improper mask generation can includes a part of the eyebrow region which may lead to wrapping artifacts and a "double eyebrow" can appear in the final image or video.

Methodology: A 19-layer VGG network [70] is implemented in [46] which is simple yet yields realistic fake results. The goal is to render the semantic content [32] of face image in the style [51] of another face image. Sometime the lighting contents of the images are not preserved in the generated images. In order to obtain the desired lighting condition of the target face, a small siamese convolutional network [30] can be constructed, which are trained using Extended Yale Face Database B [33]. This dataset contains grayscale portraits of subjects under 9 poses and 64 lighting conditions.

RSGAN [60] extracts the latent-space representation of face and hair of the input image separately using two conditional variational auto-encoder networks [45]. A GAN network called composer is used to reconstruct the face from the two latent space representations. A new conditional variational auto-encoder called DF-VAE [40] is developed to generate high quality and scalable videos. It addressed the problem of face style mismatch caused by appearance variations by introducing MAdaIN based on the original AdaIN [36]. It also addressed the temporal continuity of generated videos by using FlowNet 2.0 [37].

As the number of layers in the encoder-decoder increases, the visual quality of the synthesize faces increases but it did cost some computational time [53]. Applying a color transfer algorithm [67] during the training process drastically increases the matching of color between the source and target faces which partially solve the problem of mismatching face color. Some Deepfake techniques also use attention mask to fuse the target face into the source face [85]. Using attention mask helps in handling occlusion, eliminating artifacts, and producing natural skin tone.

Postprocessing: The postprocessing concerned on fixing the imperfect parts in the Deepfake generation process. Kalman filter [79] and Gaussian blur [39] are the most commonly used post processing methods. These filters smoothen the boundary contour of the swap face and eliminates the inconsistency and flickering of the swapped face. Using a sharpening filter on the blended faces also greatly increases the quality of the final video with no extra computational cost. Some methods use poisson image blending [65] which also helps in the final blending image but it should not be applied on the entire mask as it would blend both faces and create an "average" face rather than a face that looks like the source face. Poisson blending should only be done in a small region along the edge of the mask. Finally, the blended face images can all be combined together with audio to produce the Deepfake video using ffmpeg [18].

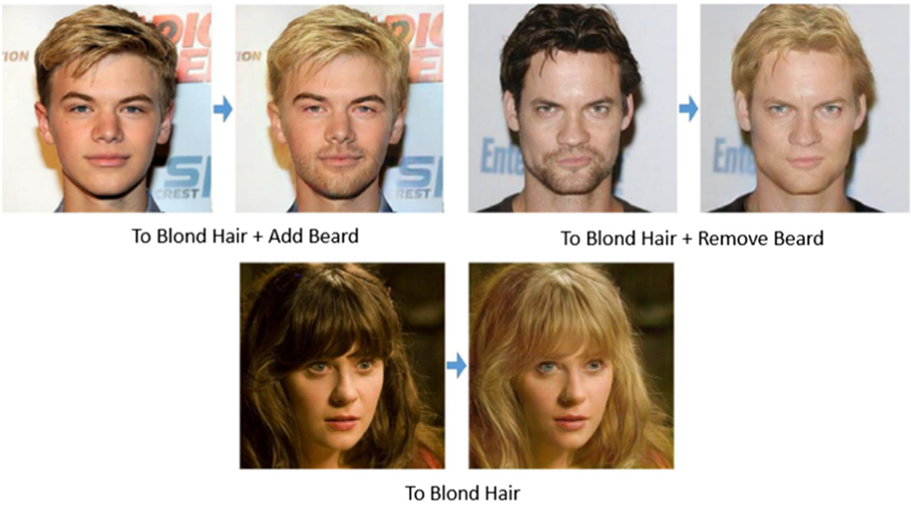

Fig. 4. Visual result of hair color manipulation for single and multiple attribute [35].

3.2 Facial Attribute Editing

Facial attribute editing is the manipulation of facial attributes (like hair color, age, gender, etc.) while preserving the identity of the person. The manipulation is performed in the attribute space. Given a face image with its corresponding attributes annotation, the editing method will give an image with altered attributes while keeping the unwanted attributes same as the original image. For developing this type of attribute editing method, celebA [56], CelebAHQ [41] and CelebAMask-HQ [50] datasets are required since the dataset are diverse with different attributes and well label annotations. Some of the examples includes changing of hair color (black, brown and blond) in Fig. 4, attribute transfer for smiling face in Fig. 5, style transfer in Fig. 6 and interactive mask modification in Fig. 7 (Table 4).

Current Editing Methods. The Invertible Conditional GANs (IcGANs) [64] are developed for complex facial image modification. This approachconsists of

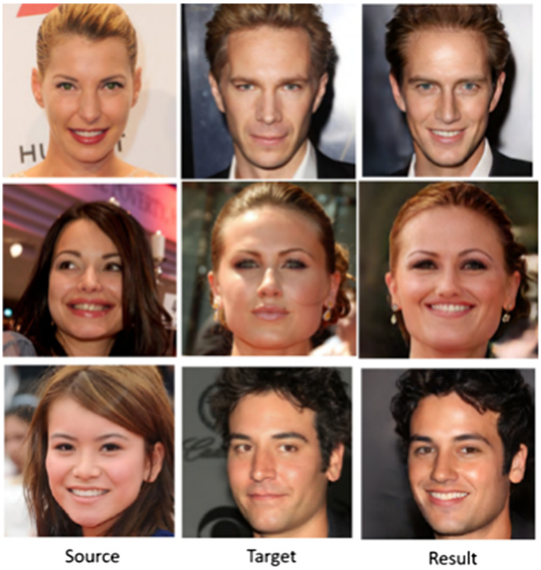

Fig. 5. Visual result of smile transfer [50].

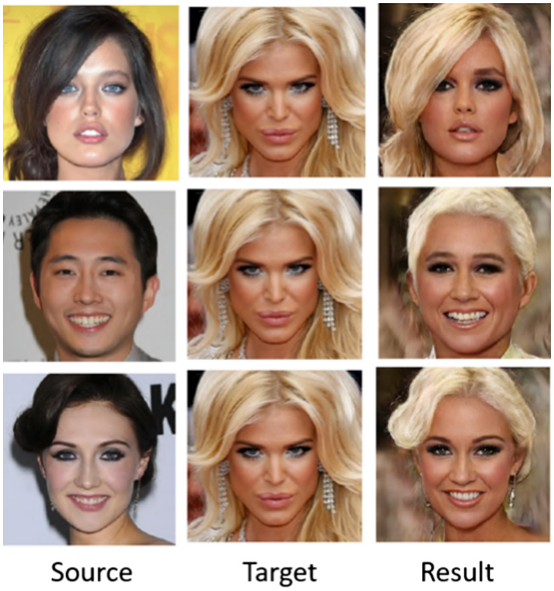

Fig. 6. Visual result of style transfer [50].

Table 3. Different real and fake dataset comparison.

Generation	Database	Year	Real data			DeepFake data			Frame/video quality	Temporal constraint	Control environment
			#Real		Origin	#Fake		Generation technique			
			#videos	#frames		#videos	#frames				
1st	UADFV	2018	49	17.3k	Internet	49	17.3k	Fakeapp	64 × 64, 128 × 128	-	-
	DF TIMIT- LQ DF TIMIT- HQ	2018	320	34.0k 34.0k	Internet	320	34.0k 34.0k	faceSwap-GAN faceSwap-GAN	64 × 64, 128 × 128	-	-
	Face forensics++	2019	1000	509.9k	Internet	1000	509.9k	Faceswap	480p, 720p, 1080p	-	-
2nd	DFD	2019	363	315.4k	Actors	3068	2242.7k	Deep fake	1080p	-	-
	CelebDF	2019	590	225.4k	Internet	5639	2116.8k	Deepake	various	-	-
	WildDeepfake	2020	3.8k	620k	Internet	3.5k	560k	Unknown	HQ	-	-
3rd	DFDC	2019	1131	488.4k	28 Actors	4119	1783.3k	Unknown	240p–2160p	-	-
	DeeperForensics 1.0	2020	50k	12.6M	100 Actors	10k	5M	Faceswap	1080p	✓	✓

Table 4. Comparation between different face attribute editing methods.

Method	Year	Dataset			# of Attributes /classes	Image scale	Architecture	Evaluation/quality assessment
		Training	Testing	Total				
IcGAN	2016	MNIST – 60k CelebA – 182k	MNIST – 10k CelebA – 20k	MNIST - 60K CelebA – 202k	0-9 (Digit) 18 (CelebA)	28 X 28 (MNIST) 64 x 64 (CelebA)	Encoder with a cGAN generator	Root mean square deviation Mean F1 score
Fader Network	2017	CelebA – 182k	CelebA – 20k	CelebA – 202k	18 (CelebA)	256 x 256	Encoder-decoder with adversarial component	Amazon Mechanical Turk (AMT) root-mean-square deviation
starGAN	2018	RaFD - 4.3k celebA – 200k	RaFD – 0.5k CelebA – 2k	RaFD – 4.8k CelebA – 202k	7 (RaFD) 3 (CelebA)	128 x 128	GAN with conditional domain information	Amazon Mechanical Turk (AMT)
AttGAN	2019	CelebA – 182k	CelebA – 20k	CelebA – 202k	13 (CelebA)	384 x 384	Encoder and decoded with attribute classifier and a discriminator	Facial Attribute Editing Accuracy/Error
STGAN	2019	CelebA – 201k	CelebA -1k	CelebA – 202k	13 (CelebA)	128 x 128 384 x 384	Encoder decoder with selective transfer unit	Facial Attribute Editing Accuracy/Error SSIM
RelGAN	2019	celebA – 200k CelebA-HQ - 27k	CelebA – 2k CelebA-HQ – 3k FFHQ - 70k	CelebA – 202k CelebA-HQ – 30k FFHQ - 70k	9 (CelebA) 9 (CelebA-HQ) 17 (CelebA-HQ)	256 x 256	A generator and three discriminator	Frechet Inception Distance (FID) SSIM user study
ResAttr GAN	2019	CelebA – 182k	CelebA – 20k	CelebA – 202k	11 (CelebA)	128 x 128 256 256	Encoder decoder generator and a Siamese network discriminator	FID SSIM Attribute editing accuracy
ClsGAN	2020	celebA – 200k	celebA -2k	CelebA – 202k	13 (CelebA)	128 × 128	A generator (two encoders and a Tr-resnet) and a discrimator	FID SSIM

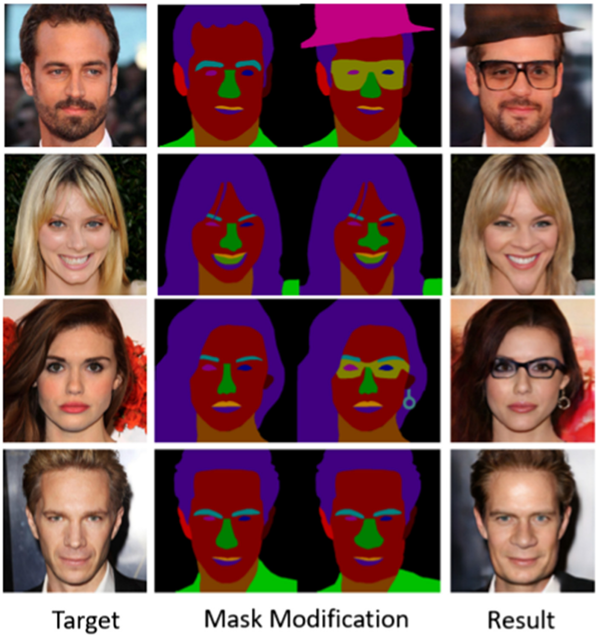

Fig. 7. Visual results of interactive face editing [50].

an encoder and followed by a conditional GAN (cGAN) [58] generator. The real image is encoded into a latent representation with the attribute information and apply variations on it to generate a new modified image. Manipulation of 18 facial attributes are chosen for this approach. The result of this approach provides face attribute changes in complex face dataset which are satisfactory. It in fact changes the identity of the person.

Fader network [48] is a new encoder-decoder architecture which is trained to reconstruct an image by disentangling the salient information of the image and the values of attributes directly in the latent space. The approach performs adversarial training for the latent space instead of the output and adversarial training aims at learning invariance to attributes. This approach makes subtle changes to portraits that sufficiently alter the perceived value of attributes while preserving the natural aspect of the image and the identity of the person. It also provides the ability to swap multiple attributes at once.

Along with development of image-to-image translation [38,82], a novel and scalable approach that can perform image-to-image translations for multiple domains using only a single model called StarGAN [29] is introduced. This architecture allows training of multiple dataset with different domains using a single network. A flexible modification of facial attributes as well as synthesizing different expressions is also observed. The attribute translations network is trained via domain classification loss and cycle consistency. This model is scalable in terms of the number of parameters required. The visual results are good and provide multiple attribute transfers. It also provides the visual comparison of the model trained in a single dataset and combination of two datasets.

AttGAN [35] is another novel approach of face attribute editing. This method removes the strict attribute independent constraint from the latent representation which was used in the previous methods, and applies the attribute-classification constraint to the generated image to guarantee the correct change of the attributes. Imposing constraints on the latent space may result in the loss of information which eventually affects the attribute editing for the worst. This approach allows only the attribute desired to change in the portrait while preserving the attribute excluding details. The results are much better for single and multiple attribute editing with higher resolutions as compared to the previous approaches.

Ming Liu et al. proposed STGAN [54] which addresses the bottleneck layers of encoder-decoder in the end giving blurry and low quality editing results observed in starGAN [29] and attGAN [35]. The selective transfer units (STUs) are integrated with encoder-decoder to adaptively select and modify encoder feature for enhanced attribute editing. This approach preserves more information of the source image as only the attributes are to be changed instead of full target attribute vector. It also proves that the difference in the attributes of the source and the target provides valuable information. The Experimental results show that STGAN simultaneously improves single and multiple attribute manipulation accuracy with higher image quality as well as perception quality. It preserves fine details and identity of source image while performing against previous methods in arbitrary facial attribute editing translation.

With an increase in the popularity of multi domain image to image translation, many limitations are being identified and fixed. The previous methods assume binary valued attributes and required to specify the entire set of target attributes, even if most of the attributes would not be changed which eventually affects the result of the manipulation. Po Wei Wu et al. introduced relative-attribute-based method, dubbed RelGAN [80] to address these limitations. The key idea is to use relative attributes which specifies the desired changes in the selective attributes. The authors proposed a matching-aware discriminator that determines whether an input-output pair matches the relative attributes and also an interpolation discriminator for improving interpolation quality. This model achieves superior performance over the state-of-the-art methods in terms of both visual quality and interpolation for the single and multi-attribute editing.

Rentuo Tao et al. proposed ResAttr-GAN [73]. Compared to existing models that perform attributes editing based on an attributes classifier, he proposed deep residual attributes learning model utilized relatively weaker information of attribute differences for face image translation. The authors proposed a Siamese-Network based residual attributes learning model to learn the attributes difference in the high-level latent space. Several facial attributes edit experiments were conducted including comparative single-attribute editing, multiple attributes editing, attributes editing on higher resolution face images, to evaluate the effectiveness of the proposed model qualitatively and quantitatively. It also demonstrated that when the training data was reduced, the proposed deep residual learning model can improve the data utilization efficiency and thus boosting the editing performance. They showed the accuracy of the attribute editing results with five different data usage. The experimental results demonstrated the effectiveness of the proposed method in both single and multiple attributes editing.

The novel ClsGAN [55] shows significant improvement in realistic image generation with accurate attribute transfer. The method introduced two approaches: Upper convolution residual network (Tr-resnet) and an attribute adversarial classifier (Atta-cls). Tr-resnet is used to extract selective information from the source image and target label. The information is acquired for combining the input and output of upper convolution residual blocks, leading to high-quality image generation with accurate attribute editing. Atta-cls is developed to improve the attribute transfer accuracy by learning the transfer defects in the generator. The comparative result shows that this method outperforms the other state-of-the-art approaches.

The previous face attribute editing is done on a predefined set of face attributes and provides very less interactive manipulation for the user. These problem are addressed in this study with the introduction of MaskGAN [50] provides a diverse and interactive face manipulation. There are two main components: Dense Mapping Network (DMN) and Editing Behavior Simulated Training (EBST). DMN learns the style mapping between the interactive mask and the target image. The architectural backbone of DMN adopts pix2pixHD [38]. Spatial-Aware Style Encoder Network inside DMN receives the style information for target image and its corresponding spacial information from the target mask at the same time. Spatial Feature Transform (SFT) from SFT-GAN [77] is used for fusing the two domains. EBST models the user editing behavior on the source mask. EBST is the overall framework composing of DMN, MaskVAE and Alpha blender, all trained together. MaskVAE is responsible for modeling the manifold of structure priors and Alpha Blender is responsible for maintaining manipulation consistency.

3.3 Face Reenactment

A Deepfake reenactment is where the source is used to manipulate the expression of the target face. It can be the manipulation of mouth, head pose, gaze, and eye blinks. Manipulation expression can provide a wide range of flexibility. Mouth reenactment, also known as “dubbing”, is when the target mouth is moved according to the movement of the source mouth. Another type is the mouth reenactment based on the audio input containing speech. Gaze reenactment is the change in the direction of target eyelids according to the movement of source eyelids. Head pose reenactment is changing of target head pose according to the source head pose. Figure 8 shows some of the reenactment approaches.

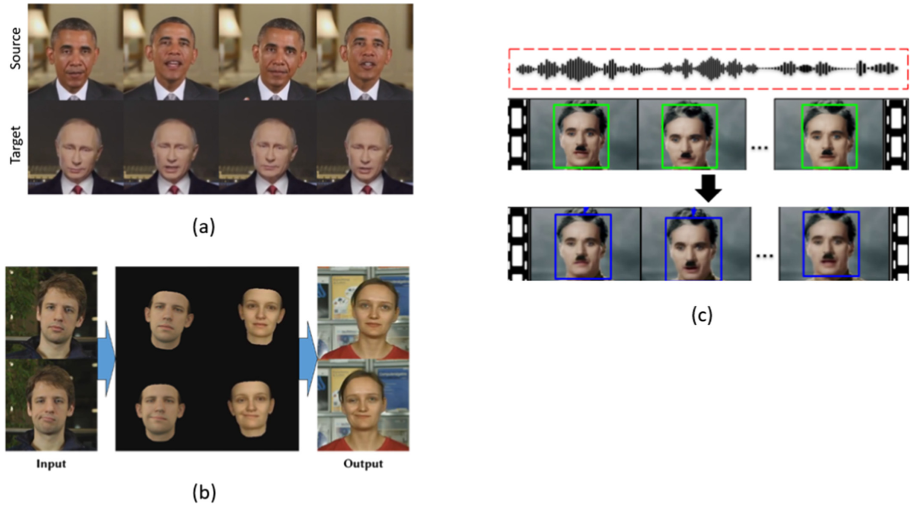

Fig. 8. Different types of face reenactment. (a) and (b) shows the full facial expression, head pose and eye motion transfer from the source to the target face with high level of photorealism [43]. (c) shows the lip-synchronization with a source and a audio segment resulting to a realistic video dubbing [66].

Current Methods. Face2Face [75] is a computer graphics based facial reenactment system that transfers the expression of a source video to a target video while maintaining the identity of the target person. This method is fully automated creating while performing face reenactment. The first frames of each video were used to obtain a temporary face identity (i.e., a 3D model), and track the expression over the remaining frames. The reenactment is done by transferring the source expression parameters of each frames of the target video.

Hyeongwoo kim et al. [43] introduced a novel approach that enables photo-realistic re-animation of portrait videos using only an input video. The first implementation to transfer full head which includes 3d head position, head rotation, face expression, eye gaze and eye blinking. This is all made possible by designing a novel GAN based space-time architecture. The network takes as input synthetic renderings of a parametric face model, based on which it predicts photo-realistic video frames for a given target actor. In order to enable source-to-target video re-animation, a synthetic target video with the reconstructed head animation parameters from a source video is rendered, and feed it into the trained network thus taking full control of the target. The author also show how we can rewrite application by combining source and target parameters. The method provides high-fidelity visual dubbing.

Wav2lip [66] is the method of lip-syncing unconstrained videos. This method produces a talking face video of an identity to match with the target speech segment. The result produces a realistic lip movement on a static image or video of an arbitrary identity. Given a short audio segment and a random reference of face image or video, the proposed model task is to generate a lip-synced version of the input that coordinates with the given audio. A quantity evaluation on the challenging benchmarks shows this method produces accurate lip-sync video which is good as real synced videos. Many application areas for this method were also discussed such as dubbing movies, translated lectures and press conferences, generation of missing video call segments, and lip-sync animations.

3.4 Face Synthesis

Face synthesis is when a new Deepfake face is created by mixing styles from different faces. It is used to produce fake persona online. Style is generally, not manually designed by a user, but extracted from reference images. Figure 9 shows the visual result of synthetic face. Two sets of image (source A and B) were generated from their latent space and the rest of the images were generated by copying a specific style from source B and taking the rest from source A. The example result shows the copying of different style corresponding to different spacial resolutions. The first row corresponds the higher level of aspect such as pose, face shape whereas the second row shows the middle level such as facial features, eye open/close, hair color and finally the last row corresponds the fine level such as eye, lighting conditions.

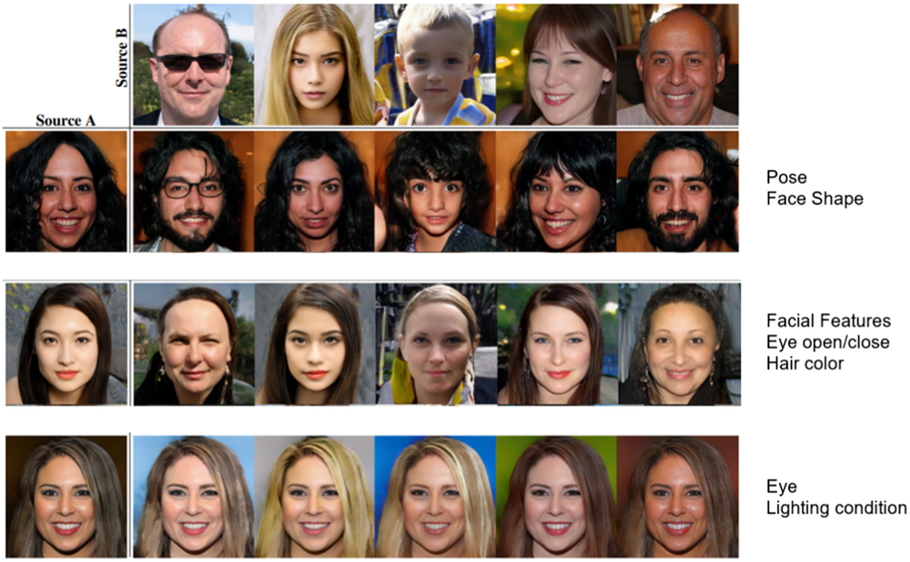

Fig. 9. Visual result of synthetic face [42]

Current Methods. A new progressive training of GAN, call PGAN [41] is introduced. The key insight is that both the generator and discriminator progressively, starting from easier low-resolution images, add new layers that introduce higher-resolution details as the training progresses. The training starts with having as low as 4×4 pixels and incrementally adds layer throughout the process till 1024×1024 pixels. The model not only generate new realistic synthetic images but also generates high quality CelebA-HQ dataset with resolution of 1024×1024.

A style based generator architecture, styleGAN [42] is also introduced. The architecture is designed to learn the style constant instead of just a feed forward network. This method has better interpolation properties, and also better disentangles the latent factors of variation. They also produces a high quality dataset called FFHQ dataset [17] which is better than the previous CelebA-HQ dataset. From two source images, set of synthesis images are generated from their latent codes. The style corresponding to coarse spatial resolution is responsible for high level aspect such as pose, hair, face shape whereas the style with middle level corresponds to facial features, eye open/close, hair style and finally with fine spacial correspond to eye, hair, lighting condition.

Semantic region-adaptive normalization, SEAN [83] is introduced. SEAN is constructed on SPADE with a generator SEAN ResNET blocks. These blocks describes the semantic regions of the segmented mask. Using SEAN normalization, a network architecture can be built that can control the style of each semantic region individually, e.g., we can specify one style reference image per region. Addition to style interpolation (bottom row), one style can be changed by selecting different styles per ResBlk.

3.5 Evaluation Methods

Root Mean Square Error (RMSE). The RMSE represents the cumulative squared error between the manipulated image and the original image.it is calculated in the pixel level. The result is better when RMSE is low.

Amazon Mechanical Turk (MTurk) [4]**.** Amazon Mechanical Turk (MTurk) is a crowdsourcing marketplace that makes it easier for individuals and businesses to outsource their processes and jobs to a distributed workforce who can perform these tasks virtually. Even when the computing technology continues to improve, there are many things that people do much more effectively than computers. Every day workers on Amazon Mechanical Turk (MTurk) help requester solve a range of data processing, analysis, and content moderation challenges. The range of includes Image and video processing, Data Verification and Cleanup, Information Gathering Data Processing. There are two ways for evaluating Deepfake: (i) one way where two images are provided in which one is fake and the other real and the user has to identify the real and the fake images and (ii) one image is provided, and user has to tell whether it is real or fake.

Structural Similarity (SSIM [78]). SSIM is an image quality assessment method based on the degradation of the structural information. It provides the similarity measurement between two images. The value ranges from 0 to 1 where higher value represents more similarities between the two images. The SSIM index can be viewed as a quality measure of one of the images being compared, provided that the other image is regarded as one of perfect quality.

Frechet Inception Distance (FID [57]). The 'Fréchet Inception Distance' (FID) captures the similarity of generated images to real ones better than the Inception Score. FID is supposed to improve the Inception Score [28] by comparing the statistics of generated samples to real samples, instead of evaluating generated samples in a vacuum. Lower FID is better, corresponding to closer distance between the generated and real data distributions.

User Study. User study is mainly the visual analysis on both real and fake images. The users are instructed to choose the best result which (i) which changes the attribute more successfully, (ii) which is of higher image quality and (iii) which one preserves the identity better with fine details of source image.

3.6 Real World Scenario

Advantages. The common benefits of Deepfake technology is their creative and productive applications in areas of education, marketing, art and multimedia [3]. Realistic video dubbing in different languages, is the next phase of dubbing. It provides the viewer an immersive experience and will think as if the person is actually speaking another language. [23] shows David Beckham lips movement with respect to the recorded audio. Digital de-aging makes a person look more younger or older digitally. It can be applied to movies, photographs, etc. In the movie "The Irishram", Robert De Niro is made to look younger [20]. Digital resurrection of a dead family member, a close friend or even a historical figure is another Deepfake application. Reanimation will bring memories or digitally interaction of a dead person. Reanimation of Salvador Dalí in the museum while taking selfies with the visitors [8]. A whole new online shopping with virtual trying on clothes or accessories [22]. Anyone can become a super model just by changing faces with your preferred body type [19]. It can also help in finding of missing children by generating faces of the appropriate age using childhood photos [6].

Drawbacks. The creation of fake porn videos, which can be used for blackmailing and taking revenge [11]. The abused of the technology can be seen when the researchers at the University of Washington posted Deepfake of President Barack Obama and spread on the internet [5]. They were able to make President Obama say whatever they want him to say. Similarly, Jordan Peele ventriloquizes Obama [24]. The fake video of Facebook CEO Mark Zuckerberg declaring "whoever controls the data, controls the future" went viral [9]. It could be a threat to the world security, increase xenophobia, violation of privacy, conspiracy theories, scams and frauds.

4 Discussion

It is becoming difficult to trust social media content as it can be fake content. With the advancement of Artificial Intelligence, realistic forensic data can be easily made. This information is misleading and it can cause distress to a lot of people in the form of hate speech, disinformation and could also stimulate political tension, public violence, or war. Deepfake can cover a large domain such as face swap and reenactment, fake news, fake photographs, fake voice, fake satellite images, and many more. Even though there are a lot of misused or unethical approaches, there is also the good side of Deepfake as was discussed before. As Artificial Intelligence can be scary sometimes, it can also be used to overcome the drawbacks brought by it. AI will be beaten by AI.

Challenges: There is always a challenge during the production of manipulated face images. The following are some of the challenges face during the production of the creation and detection process of Deepfake.

Data Necessity: For creating face Deepfake, a huge amount of data is required for the GAN to train their network. In the early stage, the number of fake datasets was in thousands but with time more fake datasets are being produced in hundreds of thousands in just a span of just 2 years. More realistic fake images are being produced in high quality but it still is not enough. A large amount of high quality real and fake images are required in the making of a good Deepfake detector system.

Data Variation: The variations in the available data are limited. There is less Deepfake dataset with eye blink or which mimics blink because nobody shares images with closing eyes. It's challenging but possible when Deepfake is generated from video extracted images where natural blinks occur. It is time-consuming when we don't get the desired images which we required for the dataset. Diverse data with distortions and noise is also needed to simulate real-world scenarios. Some of the distortion or noise can be compression techniques, blurring, and contrast change. Another diversity is how different people look and there is a need for diverse manipulated faces in terms of skin color, hair color and style, face shape, facial features, and lighting conditions. The diversity on the manipulated images or videos makes sure the data will be useful for developing a robust face forgery detection and more face related research.

Cost: Training a generative adversarial network for Deepfake is costly as it requires a lot of time to produce realistic manipulated images. There is a lot of preprocessing required before the training starts and a lot of post processing needed after. It usually takes weeks to do produce a Deepfake that looks authentic to human eyes. In order to produce fast manipulated images, high-end graphics cards are needed for training and swapping. Most of the generated manipulated image resolutions are low, so to get high-quality images, a stronger computer system will be needed.

Mandatory Post-processing: Current neural network technologies are great while producing manipulated or swapped faces but the resulting outputs are not flawless. When there is a face-swapping process, there are still some artifacts produce in the manipulated regions or on the edges. There is also a color mismatch between the swapped face and the target original face skin tone. For the current technologies, post-processing is needed to produce a final realistic manipulated face image.

5 Conclusion

This study provides the trending technology of Deepfake face manipulation especially the creation of the Deepfake. We focus on the swapping of face, editing facial attributes, face reenactment, and synthesizing a face. We also discussed the available database used in the creation of Deepfake. Some real-world scenarios were also discussed, with respect to the benefits and drawbacks of Deepfake, so not all Deepfakes are malicious.

Acknowledgment. This study was supported by the BK21 FOUR project (AI-driven Convergence Software Education Research Program) funded by the Ministry of Education, School of Computer Science and Engineering, Kyungpook National University, Korea (4199990214394).

References

1. 100k faces. https://generated.photos/. Accessed 14 Jan 2021
2. 100k generated faces. https://github.com/NVlabs/stylegan. Accessed 14 Jan 2021
3. AI enable deepfake. https://www.forbes.com/sites/bernardmarr/2019/07/22/the-best-and-scariest-examples-of-ai-enabled-deepfakes/?sh=86672662eaf1. Accessed 14 Jan 2021
4. Amazon Mechanical Turk. https://www.mturk.com/. Accessed 14 Jan 2021
5. BBC Obama. https://www.bbc.com/news/av/technology-40598465. Accessed 15 Jan 2021
6. Computer-generated age progression photos. https://www.reddit.com/r/interestingasfuck/comments/kxf12x/the_accuracy_of_computergenerated_age_progression/. Accessed 15 Jan 2021
7. Deep learning for detecting audiovisual fakes. https://sites.google.com/view/audiovisualfakes-icml2019/. Accessed 14 Jan 2021
8. Deepfake Salvador the Verge. https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum. Accessed 14 Jan 2021

9. Deepfake video of Mark Zuckerberg. https://finance.yahoo.com/news/deepfake-video-mark-zuckerberg-goes-163128674.html?guccounter=1. Accessed 15 Jan 2021
10. Deepfaketimit. https://www.idiap.ch/dataset/deepfaketimit. Accessed 14 Jan 2021
11. Deepnude. https://www.vox.com/2019/6/27/18761639/ai-deepfake-deepnude-app-nude-women-porn. Accessed 15 Jan 2021
12. Dimentions. https://app.dimensions.ai/. Accessed 15 Jan 2021
13. Facebook AI deepFake detection challenge dataset. https://ai.facebook.com/datasets/dfdc/. Accessed 14 Jan 2021
14. Faceswap. https://github.com/deepfakes/faceswap. Accessed 14 Jan 2021
15. Faceswap. https://github.com/MarekKowalski/FaceSwap/. Accessed 14 Jan 2021
16. Fakeapp. https://fakeapp.softonic.com/. Accessed 14 Jan 2021
17. FFHQ dataset. https://github.com/NVlabs/ffhq-dataset. Accessed 14 Jan 2021
18. FFmpeg. https://ffmpeg.org/. Accessed 14 Jan 2021
19. Forbes digital doubles. https://www.forbes.com/sites/katiebaron/2019/07/29/digital-doubles-the-deepfake-tech-nourishing-new-wave-retail/?sh=4e656bac4cc7/. Accessed 14 Jan 2021
20. Making Robert de Niro in "the Irishman". https://www.businessinsider.com/deepfake-netflix-correcting-the-irishman-de-ageing-tech-2020-1. Accessed 14 Jan 2021
21. NIST media forensics challenge 2018. https://www.nist.gov/itl/iad/mig/media-forensics-challenge-2018. Accessed 14 Jan 2021
22. Retailwire. https://retailwire.com/discussion/can-deepfake-technology-reduce-retail-returns-without-rattling-reality/. Accessed 14 Jan 2021
23. Reuters David Beckham's 'deep fake' malaria awareness video. https://mobile.reuters.com/video/watch/david-beckhams-deep-fake-malaria-awarene-id536254167?chan=c1tal5kh. Accessed 15 Jan 2021
24. The Verge Barack Obama. https://www.theverge.com/tldr/2018/4/17/17247334/ai-fake-news-video-barack-obama-jordan-peele-buzzfeed. Accessed 15 Jan 2021
25. Vice. https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn. Accessed 15 Jan 2021
26. Vidtimit. https://conradsanderson.id.au/vidtimit/. Accessed 14 Jan 2021
27. Workshop on media forensics. https://sites.google.com/view/mediaforensics2019. Accessed 14 Jan 2021
28. Barratt, S., Sharma, R.: A note on the inception score. arXiv preprint arXiv:1801.01973 (2018)
29. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8789–8797 (2018)
30. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), vol. 1, pp. 539–546. IEEE (2005)
31. Dolhansky, B., et al.: The deepfake detection challenge dataset. arXiv preprint arXiv:2006.07397 (2020)
32. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2414–2423 (2016)

33. Georghiades, A.S., Belhumeur, P.N., Kriegman, D.J.: From few to many: illumination cone models for face recognition under variable lighting and pose. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 643–660 (2001)
34. Goodfellow, I., et al.: Generative adversarial nets. Adv. Neural Inf. Process. Syst. **27**, 2672–2680 (2014)
35. He, Z., Zuo, W., Kan, M., Shan, S., Chen, X.: AttGAN: facial attribute editing by only changing what you want. IEEE Trans. Image Process. **28**(11), 5464–5478 (2019)
36. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1501–1510 (2017)
37. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: FlowNet 2.0: evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2462–2470 (2017)
38. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
39. Ito, K., Xiong, K.: Gaussian filters for nonlinear filtering problems. IEEE Trans. Autom. Control **45**(5), 910–927 (2000)
40. Jiang, L., Li, R., Wu, W., Qian, C., Loy, C.C.: Deeperforensics-1.0: a large-scale dataset for real-world face forgery detection. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2886–2895. IEEE (2020)
41. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)
42. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)
43. Kim, H., et al.: Deep video portraits. ACM Trans. Graph. (TOG) **37**(4), 1–14 (2018)
44. King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
45. Kingma, D.P., Mohamed, S., Jimenez Rezende, D., Welling, M.: Semi-supervised learning with deep generative models. Adv. Neural Inf. Process. Syst. **27**, 3581–3589 (2014)
46. Korshunov, P., Marcel, S.: Deepfakes: a new threat to face recognition? Assessment and detection. arXiv preprint arXiv:1812.08685 (2018)
47. Korshunova, I., Shi, W., Dambre, J., Theis, L.: Fast face-swap using convolutional neural networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3677–3685 (2017)
48. Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., Ranzato, M.: Fader networks: manipulating images by sliding attributes. In: Advances in Neural Information Processing Systems, pp. 5967–5976 (2017)
49. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T., Van Knippenberg, A.: Presentation and validation of the Radboud faces database. Cogn. Emot. **24**(8), 1377–1388 (2010)
50. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: towards diverse and interactive facial image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5549–5558 (2020)

51. Li, C., Wand, M.: Combining Markov random fields and convolutional neural networks for image synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2479–2486 (2016)
52. Li, Y., Chang, M.C., Lyu, S.: In Ictu Oculi: exposing AI generated fake face videos by detecting eye blinking. arXiv preprint arXiv:1806.02877 (2018)
53. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-DF: a large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3207–3216 (2020)
54. Liu, M., et al.: StGAN: a unified selective transfer network for arbitrary image attribute editing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3673–3682 (2019)
55. Liu, Y., Fan, H., Ni, F., Xiang, J.: ClsGAN: selective attribute editing model based on classification adversarial network. Neural Netw. **133**, 220–228 (2017)
56. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3730–3738 (2015)
57. Mathiasen, A., Hvilshøj, F.: Fast fr\'echet inception distance. arXiv preprint arXiv:2009.14075 (2020)
58. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
59. Natsume, R., Yatagawa, T., Morishima, S.: FsNet: an identity-aware generative model for image-based face swapping. In: Jawahar, C., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11366, pp. 117–132. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-20876-9_8
60. Natsume, R., Yatagawa, T., Morishima, S.: RsGAN: face swapping and editing using face and hair representation in latent spaces. arXiv preprint arXiv:1804.03447 (2018)
61. Neves, J.C., Tolosana, R., Vera-Rodriguez, R., Lopes, V., Proença, H., Fierrez, J.: GANPrintr: improved fakes and evaluation of the state of the art in face manipulation detection. arXiv preprint arXiv:1911.05351 (2019)
62. Nirkin, Y., Masi, I., Tuan, A.T., Hassner, T., Medioni, G.: On face segmentation, face swapping, and face perception. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 98–105. IEEE (2018)
63. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3D face model for pose and illumination invariant face recognition. In: 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 296–301. IEEE (2009)
64. Perarnau, G., Van De Weijer, J., Raducanu, B., Álvarez, J.M.: Invertible conditional GANs for image editing. arXiv preprint arXiv:1611.06355 (2016)
65. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. In: ACM SIGGRAPH 2003 Papers, pp. 313–318 (2003)
66. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 484–492 (2020)
67. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Comput. Graph. Appl. **21**(5), 34–41 (2001)
68. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics: a large-scale video dataset for forgery detection in human faces. arXiv preprint arXiv:1803.09179 (2018)

69. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1–11 (2019)
70. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
71. Solon, O.: Facial recognition's 'dirty little secret': millions of online photos scraped without consent. NBC News (2019)
72. Stehouwer, J., Dang, H., Liu, F., Liu, X., Jain, A.: On the detection of digital face manipulation. arXiv preprint arXiv:1910.01717 (2019)
73. Tao, R., Li, Z., Tao, R., Li, B.: Resattr-GAN: unpaired deep residual attributes learning for multi-domain face image translation. IEEE Access **7**, 132594–132608 (2019)
74. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: image synthesis using neural textures. ACM Trans. Graph. (TOG) **38**(4), 1–12 (2019)
75. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: real-time face capture and reenactment of RGB videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2387–2395 (2016)
76. Tuan Tran, A., Hassner, T., Masi, I., Medioni, G.: Regressing robust and discriminative 3D morphable models with a very deep neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5163–5172 (2017)
77. Wang, X., Yu, K., Dong, C., Loy, C.C.: Recovering realistic texture in image super-resolution by deep spatial feature transform. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018
78. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
79. Welch, G., Bishop, G., et al.: An introduction to the Kalman filter (1995)
80. Wu, P.W., Lin, Y.J., Chang, C.H., Chang, E.Y., Liao, S.W.: RelGAN: multi-domain image-to-image translation via relative attributes. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5914–5922 (2019)
81. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Sig. Process. Lett. **23**(10), 1499–1503 (2016)
82. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)
83. Zhu, P., Abdal, R., Qin, Y., Wonka, P.: Sean: image synthesis with semantic region-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5104–5113 (2020)
84. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: a 3D solution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 146–155 (2016)
85. Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.G.: Wilddeepfake: a challenging real-world dataset for deepfake detection. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2382–2390 (2020)